Environmental and Resource Economics (2023) 86:673–712
https://doi.org/10.1007/s10640-023-00807-0

Stability of Efficient International Agreements on Solar Geoengineering

Irina Bakalova[1,2] · **Mariia Belaia**[3]

Accepted: 22 August 2023 / Published online: 14 September 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract
Solar geoengineering (SG) may have the potential to reduce extreme climate damages worldwide. Yet, international coordination will make the difference between success and failure in leveraging it. Using a simple game-theoretic framework, we investigate whether the stability of an efficient, self-enforcing international agreement on SG is attainable. We demonstrate that side payments from countries less vulnerable to climate change to those more vulnerable can guarantee the stability of an efficient agreement. The size of the side payments will vary within a zone of possible agreement, which will change depending on certain key assumptions. For example, assuming stronger mitigation reduces the necessary payments. Alternatively, asymmetry in national damages from SG over-provision vs. under-provision justifies larger payments; here, the welfare-optimal strategy may be deployment that makes no one worse off. We also show that an agreement may be stable without side payments if deployment costs are substantial and counter-SG is available, while a moratorium may be socially optimal if SG brings substantial global non-excludable fixed costs.

Keywords Solar geoengineering · International environmental agreements · Side payments · Free driver · Global cooperation

JEL Classification Q54 · H41 · D62 · D02 · C72 · F53

✉ Irina Bakalova
bakalova.irina@gmail.com

Mariia Belaia
mbelaia@gmu.edu

[1] Harvard University, 12 Oxford Str., Cambridge, MA 02138, USA

[2] HSE University, 20 Myasnitskya Str., Moscow, Russian Federation 101000

[3] George Mason University, 4400 University Dr., Fairfax, VA 22030, USA

1 Introduction

Solar geoengineering (SG)—also known as solar radiation management—is a potential approach for reducing the climate change impacts of increased concentrations of greenhouse gases (GHGs) in the atmosphere, essentially by counteracting the radiative forcing associated with a given level of concentrations. Such negative radiative forcing can be produced by reflecting a share of sunlight back to space. In theory, there are several ways this could be accomplished, including painting roofs white and installing mirrors in space. One technology appears to be the most promising, namely stratospheric aerosol injection (National Academies of Sciences, Engineering, and Medicine 2021). By increasing the amount of reflective aerosols in the stratosphere, some degree of global cooling could be induced (Crutzen 2006). The naturally occurring analog is an injection of sulfur into the stratosphere following volcanic eruptions, which does, in fact, cool the planet for a limited time. Stratospheric aerosol injection (henceforth referred to as Solar Geoengineering, SG) is the focus of this paper.

In the context of the threat of severe damages from climate change and the political and other challenges of enacting meaningful climate policies to reduce GHG emissions, SG emerges as a potentially useful instrument as part of the climate policy portfolio (Belaia et al. 2021). It would act fast, and its direct deployment costs have been estimated to be relatively low (Smith 2020). This potential, however, comes with significant caveats. First, while SG could be deployed by individual countries, its impacts would be global. Second, other than the impact on global temperature, a reduction in radiative forcing from the use of SG will not have the same effect on the global and regional climates as that which would occur from an equivalent reduction in radiative forcing from reducing the atmospheric concentration of GHGs by either reducing emissions or increasing removals. Third, SG will also not address related issues such as ocean acidification. Fourth, beyond a certain 'optimal' level of SG, it is expected that the negative effects of any additional SG will exceed its positive effects, and this 'optimal' level will vary by country. Accordingly, Weitzman (2015) described SG as a "public gob"; that is, it can be either a public good (i.e., has positive externalities) or a public bad (i.e., has negative externalities), and this may vary from region to region for a given level of SG.

Thus, as in the case for emissions reductions, nationally optimal SG levels will differ across countries, an issue that has been referred to as the "global thermostat problem" (Rickels et al. 2020). In both cases, therefore, international coordination can make the difference between success and failure. Global cooperation on emissions reductions is plagued by a free-rider problem—how to encourage countries to take action, i.e., reducing their own emissions, rather than simply benefitting from the positive externalities resulting from the actions of others. SG presents the opposite challenge—what Weitzman (2015) characterized as a "free-driver" problem—of how to prevent action taken in the self-interest of jurisdiction, which may result in negative externalities for others.

A key question, therefore, is whether a (self-enforcing) stable and efficient international agreement on SG can be reached. To this end, Weitzman (2015) demonstrates that when all states have a capacity to act, the country that prefers the lowest temperature will determine the temperature in the absence of cooperation. This is almost certainly an inefficient outcome, as this level of SG will be greater than optimal for all other countries. Weitzman (2015) went on to argue that, in theory, an efficient outcome could be achieved by forming a grand coalition that chooses some SG level via a voting rule. However, the question of stability remains, since countries that prefer more SG would have an incentive to free drive.

Furthermore, as pointed out by Heyen et al. (2019), if countries can counter-geoengineer, though, this could lead to a "climate clash", a mutually destructive outcome.

Ricke et al. (2013) show that assuming that countries outside of the coalition cannot deploy SG and all countries within the coalition gain from membership, there is an incentive to create the smallest exclusive-membership coalition with enough power to deploy and sustain SG. Similarly, Lloyd and Oppenheimer (2014) argue for restricted membership, despite the weak legitimacy, since smaller coalitions are more effective. In both cases, there is no proposal for a mechanism that deters non-members from deployment. Hence, the stability of the coalition is unaddressed.

Parson (2014) suggests an exclusive-membership coalition, where the ability of SG deployment is linked to emissions cuts. This might solve the mitigation under-provision problem but may require an unprecedented level of commitment from countries. That said, the potential interplay between mitigation and SG is not straightforward and the topic has attracted substantial interest among game theorists. Urpelainen (2012) demonstrates that the magnitude and asymmetries of negative externalities from SG determine whether it acts to increase/decrease mitigation. A stylized extensive study of tradeoffs between mitigation and SG in the noncooperative setting with asymmetric countries is offered by Manoussi and Xepapadeas (2017). A behavioral experiment in Cherry et al. (2022) demonstrates that the mere option to deploy SG may increase mitigation efforts. In this spirit, Millard-Ball (2012) finds that if collateral damages from SG deployment are substantial enough, countries would have an incentive to increase mitigation efforts to disincentivize potential free drivers from unilateral deployment. What's more, substantial collateral damage may foster a self-enforcing climate treaty with full participation. In the follow-up study, Finus and Furini (2023) add that collateral damages should be not too high for the unilateral deployment to be a credible threat. Using the WITCH regional integrated assessment model, Emmerling and Tavoni (2018) find that in both cooperative and non-cooperative scenarios, SG reduces optimal mitigation. Worth noting, none of the studies suggest SG may replace mitigation.

Rickels et al. (2020) explore what constitutes a globally efficient level of SG, suggesting that it is the value that minimizes gross value added (GVA)—weighted average gap between country-specific actual climate and output-maximizing climate. In their discussion, which is grounded in estimations based on economic and climate projections, they conclude that global cooperation on SG is likely. Yet, their study does not offer formal game-theoretic derivations and analysis of international agreement's stability.

More generally, SG poses several governance challenges. Some are novel, while others have historical analogies. Deterrence of unilateral unauthorized SG deployment is often compared with non-proliferation of nuclear weapons. In this spirit, political scientists (e.g., Keohane and Victor 2011; Bunn 2019; Nye 2019) have suggested that approaches for SG governance may be drawn from the experience of the "regime complex", which is characterized by a core treaty plus a set of supplementary formal accords and informal initiatives. This literature further suggests that countries would need to build up a set of norms against unilateral SG deployment.

Further analogies with the nuclear regime should be made with caution. First, SG can be a public good and a public bad, and the consequences of small amounts are less likely to be catastrophic. Second, the larger number of potential actors will make the SG negotiation process even more complex. The "gob" nature of SG and therefore the absence of mutually harmful outcomes that parties will want to avoid makes this situation fundamentally different from conflicts considered in Schelling's concept of stability (Schelling 1990), and this affects bargaining strategies. However, mutual deterrence can still be reached by

considering counter-measures to SG deployment, and this option has been explored by Parker et al. (2018) and Heyen et al. (2019).

In this paper, we offer an economic analysis of the stability of an efficient international agreement on SG by developing a simple game-theoretic framework and considering a range of model variations. Our focus is limited to SG, and therefore we do not consider potential interactions with other climate policy instruments. To address the unique challenges posed by SG, we use a hybrid approach to the stability analysis of an efficient SG agreement and the associated design of side-payments scheme, allowing SG to be both a public bad and a public good. More specifically, we combine the concept of stability of self-enforcing agreements, which is native to the context of public goods, with an adaptation of the Coase theorem, which is native to the context of public bads. Our analysis focuses on positive incentives to sustain global cooperation on SG: side payments. Import tariffs and trade restrictions have been the subject of a study by Eigruber and Wirl (2018), who highlight that unilateral deployment, followed by trade restrictions, would result not only in a decrease in global welfare but potentially increase domestic emissions in those countries that deployed SG. With this, we find side payments to be an alternative instrument worth exploring.

We demonstrate that side payments directed from countries less vulnerable to climate change toward those more vulnerable can guarantee the stability of an efficient agreement. Further, our results suggest that there is room for bargaining regarding the exact side-payments scheme. We also find that, in some cases, an agreement may be stable without side payments: in particular, when deployment costs are substantial, or counter-SG is available, recognizing that deployment costs are currently estimated to be almost negligible, and counter-SG may be a hazardous and inefficient action due to wasted resources. Finally, a moratorium on SG deployment may be key to a self-enforcing, efficient, and stable agreement with side payments when SG is associated with substantial non-excludable fixed costs.

The paper is organized as follows. The following section introduces the game-theoretic framework, Sect. 3 offers the model parametrization. Section 4 presents model results and offers a case study for alternative problem settings, investigating the implications of counter-SG, non-negligible deployment costs, mitigation, universal fixed costs associated with SG deployment, and country weights in global decision-making. Section 5 summarizes our findings and concludes. Analytical derivations are offered in the Appendix.

2 The Game-Theoretic Modelling Framework

The stability and effectiveness of environmental agreements have long been the subject of scholarly inquiry (Chander and Tulkens 1992; Carraro and Siniscalco 1993; Barrett 1994; Lessman et al. 2015; Meya et al. 2018; Finus et al. 2021). In the context of climate change, studies have largely focused on emissions abatement, characterized by the problem of under-provision. SG, in contrast, presents the problem of over-provision. In this sense, the challenge may be closer to that of agreements on nuclear non-proliferation and further from the challenges of agreements on limiting GHG emissions.

We characterize an agreement with voluntary participation as being *stable* if it is both internally and externally stable. According to the definition of d'Aspremont et al. (1983), an agreement is *internally stable* if none of the signatories wishes to withdraw, and *externally stable* if none of the non-signatories wishes to join.

It is often impossible to achieve a self-enforcing stable coalition without any additional instruments. There is a wide range of potential policy instruments that could be used, in theory, to sustain cooperation, including sanctions, trade measures, and even military intervention. Here, we focus on positive incentives, such as side payments. As shown by Barrett (2001), side payments among asymmetric in their benefits from the abatement countries may substantially improve the stability of cooperative international agreements. Regarding a free-riding problem, the internal stability of a coalition with side payments requires that the joint payoff of signatories exceeds the sum of payoffs that countries would get in the case of a unilateral deviation from cooperation. In this case, payoffs in the coalition may be distributed in a way that no one is better off by unilaterally deviating; this is known as *the optimal transfers scheme* (Carraro et al. 2006; Caparrós and Finus 2020). Technically, such a scheme may be applied to the SG game considered here, albeit it would lead to a limited set of stable coalitions as detailed in the next paragraph. Instead, we propose a side-payments scheme that achieves stability for all possible coalitions including the global cooperation in a default setting, and also for a range of alternative settings.

Unlike in the case of emissions abatement, the stability of an agreement on SG is undermined by free-driving incentives. Since the direct deployment expenses may be considered negligible, countries do not save costs when deciding to unilaterally deviate from an agreement. Rather, the reason for the deviation is to deploy an extra amount of SG to reach the nationally preferred level. As a consequence, every possible coalition has a clearly defined group of countries that do not have an incentive to deviate—these are countries with preferences for lower levels of SG. We refer to these countries as *non-drivers* and those countries that do have the incentive to deviate as *free drivers*. Therefore, the assumptions related to unilateral deviation used in the optimal transfers scheme do not hold for *non-drivers*. Our proposed side-payments scheme offers an alternative reference point for these countries with the associated game structure that reflects the unique nature of SG.

Non-drivers have the incentive to deter free drivers from SG over-provision, which can be done by means of compensating the free drivers for the higher local climate damages associated with lower levels of SG. For this, we introduce the establishment of an international fund for such compensation. This fund may be thought of as an adaptation fund since it is directed at reducing local damages from climate change. If the size of such fund is not sufficient to cover all potential compensation claims, then some free drivers would choose to deviate, and cooperation fails. Therefore, we suggest that for non-drivers the more plausible reference point is the absence of cooperation.

As we will demonstrate in Sect. 2.2, below, any SG coalition can be internally stable with side payments if the non-drivers' total willingness to pay to sustain cooperation is greater than or equal to the total willingness to accept of the free drivers to continue to cooperate.

2.1 The Model

Our modeling framework is based on seven assumptions regarding impacts, technology, and game structure:

[A-1] SG deployment capability Every country is capable of SG deployment, either unilaterally or as part of a coalition.

[A-2] Symmetry of damage function For each country, the total damages from climate change and SG implementation are symmetric around their optimal level of SG

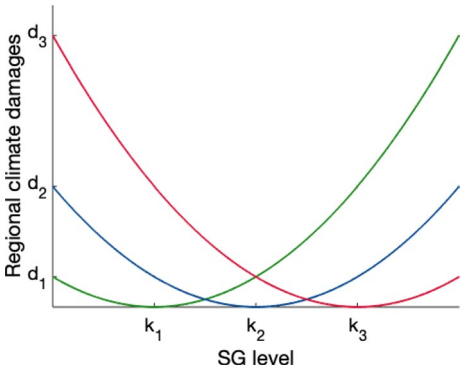

Fig. 1 National climate damages as a function of SG level, illustrated for three countries. Here, d_1, d_2, and d_3 indicate damage costs in the absence of SG; and k_1, k_2, and k_3 are the optimal SG levels for each of the three countries

(see Fig. 1). This simplifying assumption eases analytical derivations and is subsequently relaxed in the Results section.

[A-3] Negligible SG deployment costs Direct deployment costs are negligible and can, therefore, be ignored. Currently these are estimated to be about 18 billion 2020 USD per degree Celsius of warming avoided (Smith 2020). This amounts to about 0.02% of global GDP, while the benefits of deployment may exceed 1% of global GDP (Belaia et al. 2021). In addition, direct deployment costs are negligible compared with climate damages and climate policy costs.

[A-4] No counter-SG is available Under counter-solar geoengineering (CSG) we refer to technical approaches that are directed at reversing SG-driven changes in radiative forcing. In principle, CSG could be either countervailing, where something is introduced to counteract the RF effect of SG, or neutralizing, where the SG agent is removed or nullified (Parker et al. 2018). To the best of our knowledge, such technologies are speculative and are not being studied actively.

[A-5] Agreement on SG is separate from emissions mitigation agreement (s) Our focus is limited to SG and, for simplicity, we treat the level of emissions as exogenous.

[A-6] We limit our consideration to a single coalition formation, which is a common approach in the international environmental agreements' literature (Barrett 1994; Carraro et al. 2006; Diamantoudi and Sartzetakis 2006; Finus and McGinty 2019).

[A-7] Potential free drivers are not foresighted That is, in their decision-making, they assume that if one country deviates from the coalition, other countries will continue to cooperate.

In Sect. 4, we explore the implications of relaxing assumptions [A-2]–[A-4] and the implications of variations in the mitigation level. Next, we present the model design.

Let us consider a set of n countries N, $N=\{1,2\ldots n\}$, with heterogeneous climate damages, which can be reduced via SG. As such, e.g., ocean acidification is not included in this definition. Let $g_i \geq 0$ Wm^{-2} be the amount of SG deployed by an individual country $i : i = 1, 2, \ldots n$. The realized SG level, G_N, Wm^{-2}, is then given by $G_N = \sum_{i=1}^{n} g_i$.

We assume that, for each country, climate damages increase quadratically with radiative forcing. Thereby, the marginal reductions in national climate damages decrease with the amount of SG, i.e., the damages decrease at a decreasing rate. The optimal level of SG occurs when the marginal reductions are 0. Beyond this point, increasing side effects and over-compensation from SG drive climate damages back up. Potential side effects include air pollution and ozone loss, which increase with additional SG deployment. In

accordance with the simplifying assumption A-2 above, the national damage functions are symmetric around each country's optimal level of SG. As a result, the national damage D_i function (illustrated in Fig. 1) can be specified as follows:

$$D_i(G_N) = \frac{\alpha}{2}(G_N - k_i)^2, \tag{1}$$

where $\alpha > 0$ [billions USD per $(\mathrm{Wm}^{-2})^2$] defines the steepness of the damage function and is the same for all countries; $k_i \geq 0$ denotes the level of SG that is optimal (in the sense of climate damage minimization) for the country i. Here, we parametrize SG in terms of negative radiative forcing. The country i's climate damages in the absence of SG are given as follows:

$$D_i(G_N)\Big|_{G_N=0} = \frac{\alpha}{2}k_i^2 \equiv d_i.$$

The case with a more general form of the damage function with the explicit d_i is presented in the Appendix. As such, $d_i > 0.5\alpha k_i^2$ indicates the climate damage that cannot be compensated via SG. That is, it includes GHG-driven climate damages such as ocean acidification.

The game The game has three stages. In the first stage, countries decide independently whether to join a SG deployment coalition. In the second stage, countries create a fund for compensations that will be used in the third stage. All coalition members decide whether and how much to contribute to the fund. In the third stage, players simultaneously choose their deployment levels. Non-signatories choose the level of SG to minimize their national damages. Coalition members choose the deployment level by minimizing the sum of damages of all signatories and implement the side payments scheme that guarantees all potential free drivers the benefits no less than what they get from their respective unilateral deviation. However, if the size of the established fund is not sufficient to cover all compensation claims, potential free drivers would deviate and the cooperative agreement collapses. In this case, the contributions made in the second stage are fully refunded. Essentially, the second stage is a threshold public good game with fully refunded contributions.

Overall, the game setting captures the side payments scheme that addresses both free riding and free driving incentives that arise in a quest for an efficient SG deployment agreement.

To solve the game, we use backward induction. We begin by solving the climate damage minimization problem for coalition signatories and for each non-signatory under all possible coalition compositions $S \subseteq N$, in the absence of side payments. We then identify the minimum fund size necessary to compensate all potential free drivers. After that, we consider the contributions to the fund by coalition members, and finally, we analyze the stability of the coalition.

2.2 The Solution

The default model specification is solved analytically, and so our results are independent of the specific distribution of countries' optimal SG levels, k_i. Below we provide the backward induction solution of the considered game. Detailed derivations for a more general functional form are presented in "Appendix A".

Third stage: optimal deployment In the third stage, players define their SG deployment level. Individually optimal deployment level for a country i is equal to k_i. For an arbitrary coalition S, $S \subseteq N$, the sum of all members' damages (D_S) is minimized:

$$\min D_S = \frac{\alpha}{2}\Big(s\big(G_S + G_{N\setminus S}\big)^2 - 2\big(G_S + G_{N\setminus S}\big)\sum\nolimits_{i\in S} k_i + \sum\nolimits_{i\in S} k_i^2\Big),$$

where s is the number of coalition members, $G_S = \sum_{i\in S} g_i$ is the total SG deployment level by a coalition, $G_{N\setminus S} = \sum_{i\in N\setminus S} g_i$ is the total deployment of non-signatories. Since the direct deployment costs are negligible, SG deployment effort may be distributed among countries in multiple ways without any alteration to national costs or benefits. The optimal SG deployment of a coalition S results from the minimization problem and constitutes the average of optimal SG levels of all of its members:

$$G_S = \frac{\sum_{i\in S} k_i}{s} \equiv \overline{k_s}.$$

The strategy for each player (non-signatories or a coalition) is to deploy the amount of SG that meets the player's desired SG level or do nothing as there is no option to reduce deployment done by others. For each country i outside of the coalition S ($i \in N \setminus S$), the deployment strategy takes into account its own preferred SG level and the total deployment levels of all other countries. The country i reaction function reads:

$$g_i = \max(0, k_i - \sum\nolimits_{j\in N\setminus\{i\}} g_j) \text{for} i \in N \setminus S.$$

For members of a coalition, the deployment strategy takes into account the preferences of all signatories and the deployment level of non-signatories. The coalition S reaction function reads:

$$G_S = \max\Big(0, \overline{k_S} - \sum\nolimits_{j\in N\setminus S} g_j\Big) \text{for} S \subseteq N.$$

We can distinguish two types of coalitions—active and passive. An *active coalition* deploys SG and defines the global thermostat. It is characterized by the optimal SG level $\overline{k_S}$, which is greater than the preferred deployment level of non-signatories, i.e., $\overline{k_S} \geq \sum_{i\in N\setminus S} g_i$. Coalitions of the second type are *passive* in the sense that the thermostat is defined by non-signatory (-ies) with a preference for a level of SG above that of a coalition: $\overline{k_S} < \sum_{i\in N\setminus S} g_i$. The second type of cooperation is always sustainable but shows zero effectiveness. Therefore, we further focus solely on the active coalitions.

When a coalition defines a desired SG level, G_S, it divides countries into two groups:

(i) Countries with optimal SG levels above G_S (henceforth, *free drivers*); and
(ii) Countries with optimal SG levels below or equal to G_S (henceforth, *non-drivers*).

Let us denote by DR^S the set of drivers, and by ND^S the set of non-drivers of coalition S, i.e.:

$$DR^S = \{j \in S : k_j > G_S\}, ND^S = \{i \in S : k_i \leq G_S\}.$$

In the third stage of the game, free drivers have an incentive to deviate by unilaterally deploying an additional amount of SG. To measure the minimum amount that these countries are willing to accept to abstain (WTA) from free driving, i.e., a break-even point, we

estimate the reduction in a driver's climate damages associated with its unilateral deviation relative to the cooperation. Here, we *separately* analyze WTA for each free driver, which is different from the non-cooperative equilibrium case. More specifically, in the non-cooperative case, only the country h that prefers the largest SG level, minimizes its climate damages, while all other countries experience additional damages due to SG over-provision. If the coalition only deters country h from deviation, then the country with the second largest preferred SG level would deviate to reach its optimum. Therefore, this country also needs to be compensated.

Overall, the size of the side payments fund created in the second stage should be greater than or equal to the sum of benefits for all drivers from reaching their optimal point. Then the fund can be distributed via side-payments among all drivers and neutralize their incentives to deviate. Otherwise, if the size of the fund is less than this amount, drivers will deviate, the coalition will collapse to the non-cooperative setting, and all contributions will be refunded. Let us denote the minimum necessary size of the fund for a coalition S as $F_S = \sum_{j \in DR^S} \left(D_j(G_S) - D_j(k_j) \right)$.

Second stage: fund establishment In the second stage coalition members make their contributions to the fund. Since only non-drivers are interested in sustaining the cooperative SG level, only they will make contributions. To estimate the maximum amount non-drivers are willing to pay (WTP) to prevent free driving, we look at the benefits to non-drivers from cooperation relative to the non-cooperative case. Any vector of individual contributions c_i such that

$$c_i \leq D_i(k_h) - D_i(G_S) \forall i \in ND^S,$$

i.e., for each non-driver, the contribution is less than or equal to its WTP, and.

$$\sum\nolimits_{i \in ND_S} c_i = F_S,$$

I.e., the sum of contributions reaches the minimum necessary size of the fund, is a strict Nash equilibrium in this subgame. If any non-driver decreases its contribution below this amount, the threshold would not be reached. The country then loses its benefit from the cooperation that is above the contribution that it could have saved. On the other hand, by increasing the contribution beyond c_i, this country will only increase its costs without additional benefit. Therefore, if the total WTP of non-drivers is greater than or equal to F_S, which is the total WTA of free drivers, then coalition members can create a sufficiently large side-payments fund to satisfy all of the compensation claims of free drivers realized in the third stage of the game.

We can define the bargaining zone as the range between the total willingness of non-drivers to pay for deterring free driving and the total willingness of drivers to accept to abstain from free driving. When total WTP minus total WTA is negative, then the stable and efficient agreement cannot be reached (or justified with benefit–cost analysis by parties) by side payments alone. When this value is zero, there is only one possible value of side payments leading to a stable self-enforcing agreement—the break-even point for both drivers and non-drivers. Finally, when this value is positive, there exists a zone of possible agreement (ZOPA), that is, a range of possible levels of compensation leading to a stable self-enforcing agreement. Which of the possible levels of compensation is realized depends on the relative bargaining power between the free drivers and the non-drivers.

First stage: the participation decision In the first stage of the game, countries decide whether to join a coalition or not. In the following subsection we demonstrate that under

the default settings, countries are always better off as part of a coalition. This holds for all coalitions $S \subseteq N$, including global cooperation.

2.3 Stability and Effectiveness Analysis

Stability The stability of the considered SG agreement requires that coalition members make sufficient contributions to the side-payments fund to compensate for climate change damages of potential free drivers associated with under-provision of SG relative to their respective nationally preferred level. For this, the total willingness of drivers to accept to abstain from free driving should not exceed the total willingness of non-drivers to pay for deterring free driving.

For the default model setting, we state the following Proposition:

Proposition: The total gain to non-drivers from cooperation (relative to the non-cooperative equilibrium) is greater than or equal to the sum of the gains to all drivers from their unilateral SG deployment above G_S *(i.e., relative to cooperation):*

$$\sum\nolimits_{i\in ND^S}\left(D_i\left(k_h\right)-D_i\left(G_S\right)\right)-\sum\nolimits_{j\in DR^S}\left(D_j\left(G_S\right)-D_j\left(k_j\right)\right)\geq 0.$$

The proof is available in "Appendix A2". The proposition states that in our default setting, the bargaining zone is always non-negative. That is, the cooperative outcome can be sustained by a stable agreement with side payments. This is a consequence of the quadratic shape of the damage function, where for each country, marginal damages increase as SG rises above the country's optimal SG value. Note that the result can be extended to a more general functional form with positive residual damages at each country's optimal deployment point. This is demonstrated in Appendix "Appendix A2".

The proposition is in the spirit of the Coase Theorem, but there is an important difference: free drivers in a unilateral deviation from cooperation achieve the minimum of their individual damage functions, while in the non-cooperative equilibrium, all countries except for the one with the highest preferences suffer from SG over-provision (see the illustration in the "Appendix A3", Fig. 9). This means that when one free driver considers a deviation, it assumes that all other potential drivers continue to cooperate. Since we show that all partial coalitions are potentially stable with the proposed side payments, this assumption is plausible: all other potential drivers sustain the cooperative SG deployment level as their losses are sufficiently compensated by non-drivers.

Effectiveness The effectiveness of a coalition S is measured as the difference between the corresponding total damages and the total damages in the social optimum, i.e. under global cooperation, with larger values indicating less effectiveness.

Global cooperation Here, net global damages D_N^* are given by:

$$D_N^*=d_N-\frac{\alpha}{2}n\overline{k_N}^2,$$

where $d_N=\sum_{i\in N}d_i$ denotes global damages in the absence of SG. Because the second term is non-negative, then $D_N^*\leq d_N$. In other words, society overall is better off with SG deployment. This is the minimum level of global damage countries may achieve with SG deployment, which is a socially optimal outcome.

Partial cooperation An arbitrary coalition S results in the following level of global damages:

$$D_N^S = d_N - \frac{\alpha}{2} n \overline{k_S} \left(2\overline{k_N} - \overline{k_S} \right)$$

The effectiveness of a coalition S scales as the square of the difference between the average of the preferred SG deployment levels in the world and a coalition:

$$D_N^* - D_N^S = \frac{\alpha}{2} n \left(\overline{k_N} - \overline{k_S} \right)^2.$$

Therefore, the more the distribution of optimal points in a coalition S is "representative" of the distribution of optimal points for all countries, the more effective a coalition is.

Non-cooperative equilibrium In the non-cooperative equilibrium, the SG level is determined by the preference of the country with the highest level of optimal SG, k_h. This level induces considerable losses for other countries, and total losses constitute:

$$D_N^{Nash} = d_N - \alpha n \overline{k_N} k_h + \frac{\alpha}{2} n k_h^2$$

Total losses in the non-cooperative equilibrium may even exceed total climate damage costs in the absence of SG deployment. This is the case when the preferred SG level, k_h, is substantially larger than the socially optimal level. More specifically, in our setting, it would need to be more than twice as great as the socially optimal level of SG: $k_h > 2G^*$. Here, total (global) damages associated with SG deployment exceed the total benefits of compensating for GHG-driven radiative forcing. This means that in a situation characterized by an inability of countries to agree on the SG level due to a lack of international cooperation and in the presence of a country that prefers an amount of SG substantially greater than that for every other country—a moratorium on SG may be the preferred solution, at least until an agreement is reached. Yet, it does not represent an optimal solution to the underlying problem of reducing regional climate damage. That is, as we discuss below, a stable agreement on SG that both deters free drivers and leverages SG to reduce local damages is preferable to a moratorium.

Our optimistic result suggests that an efficient and stable agreement on SG is attainable by means of side payments directed from countries less vulnerable to climate change (SG preferences below G*) to those more vulnerable (SG preferences above G*), in theory. In reality, it may appear impossible to attribute/agree on the attribution of local climate impacts to SG vs. GHG-driven changes. Thus, as an alternative, a global risk side-payments scheme that allows for risk transfer without attribution may be used. As such, Horton and Keith (2019) suggest a system of multilateral parametric climate risk insurance.

3 Model Parametrization

3.1 Parametrization Approaches

Some specifications cannot be solved analytically. For these, we pursue numerical solutions using the GAMS software. For model calibration, we consider four separate distributions of SG preferences across countries. The first three are theoretical distributions that span distinct distribution shapes. The final one is based on empirical data on the regional costs and benefits of SG taken from Rickels et al. (2020). In particular, we use the data for climate and economic background conditions as given in SSP5 under the IIASA growth projections for the year 2050.

Theoretical parametrization Here, we consider seven countries and generate three alternative discrete mean-preserving distributions of optimal SG levels across countries, as follows:

i. Uniform distribution*:* $k_i \in K^U = \{1, 2, 3, 4, 5, 6, 7\}$, Wm^{-2};
ii. Left-modal distribution. The majority of countries prefer lower amounts of SG with one potential driver*:* $k_i \in K^L = \{2.6, 2.8, 3.1, 3.3, 3.5, 3.7, 9.0\}$, Wm^{-2};
iii. Right-modal distribution. The majority of countries prefer larger amounts of SG:$k_i \in K^R = \{1.0, 1.2, 4.5, 4.8, 5.3, 5.5, 5.7\}, \mathrm{Wm}^{-2}$.

In all three cases, the mean value of k_i is 4 Wm^{-2}. We set $\alpha = 1.5$ billion USD per $(\mathrm{Wm}^{-2})^2$. With this, national climate damages in the absence of SG range between 0.75 and 60 billion USD in the three cases above. This same calibration is used for all numerical simulations throughout the paper.

In the results illustration, both benefits to nondrivers from cooperation, and benefits to drivers from unilateral deployment above G*, are normalized by the no-SG global climate damages, d_N.

Empirical calibration We base our calibration on a study by Rickels et al. (2020), who estimate the change in gross value added (GVA) following the deployment of SG at a globally efficient level for 178 countries.

First, we estimate each country's optimal SG level, k_i. Let us define $V_i(0)$ and $V_i(G^*)$ as the country i's GVA in the absence of SG and at the globally optimal level of SG, respectively. We then define the normalized change in GVA following SG deployment as:

$$\frac{V_i(G^*) - V_i(0)}{V_i(0)} \equiv V_i$$

To be consistent with the definition of an efficient SG level from Rickels et al. (2020), we assume that G^* is the GDP-weighted average preferred SG level among countries. We then find k_i by solving the system of simultaneous equations,[1] for all 178 countries:

$$\frac{2}{\alpha}(D_i(0) - D_i(G^*)) = 2G^* k_i - G^{*2} = V_i, \forall i \in N.$$

Then we put a lower bound of 0 on k_i and proportionally adjust the preferences of countries such that the globally efficient level is $4\mathrm{Wm}^{-2}$, following Rickels et al. (2020).

Finally, to reduce the computational cost, we limit our consideration to 19 out of 178 countries in the dataset. We chose countries that benefit or lose the most (in absolute terms) from globally efficient SG, treating the European Union as one decision-making unit.[2] As a result, we arrive at the 19 countries listed in Table 1. Together, these countries amount to 75% of the change in global GVA following the deployment of SG at a globally efficient

[1] The dataset in Rickels et al. (2020) allows to calibrate k_i but not α. In an attempt to avoid merging distinct and potentially conflicting climate-impacts datasets, we set α at a constant value across countries. This simplifying assumption implies that national damage functions are calibrated so that damages in the absence of SG reflect relative vulnerability to climate change across countries but do not necessarily depict an accurate associated level of damages.

[2] We chose the countries with benefits or costs above 1% of the change in global GVA.

Table 1 List of countries that benefit/lose the most from the socially optimal level of SG (relative to their GVA in the absence of SG), their preferred SG level, and their power index in 2017

Country	Abbreviation	Optimal SG level (W/m^2)	World power index 2017
Venezuela	VEN	8.39	0.575
Saudi Arabia	SAU	7.68	0.733
Bangladesh	BGD	6.9	0.491
Nigeria	NGA	6.57	0.504
Congo	COD	6.52	0.342
Brazil	BRA	6.36	0.74
Indonesia	IDN	6.03	0.645
Egypt	EGY	5.64	0.569
India	IND	4.88	0.707
Pakistan	PAK	4.09	0.549
Mexico	MEX	4.09	0.695
Australia	AUS	3.04	0.788
Japan	JPN	1.13	0.851
China	CHN	0.71	0.861
USA	USA	0.52	0.954
EU	EU	0.49	0.841[a]
The UK	GBR	0.00	0.817
Canada	CAN	0.00	0.8
Russia	RUS	0.00	0.758

[a]For EU, we use the largest index value among EU members, which is the one for Germany

level. The associated distribution of preferences over SG level across countries is given in Table 1.

It is worth highlighting, following Rickels et al. (2020), that our calibration reflects that some countries may benefit from climate change and, as a result, any level of SG would bring them a net decrease in GVA. Among the selected 19 countries these are Canada, Russia, and the UK. This is in contrast to other game-theoretic studies that assume that some amounts of SG benefit all countries (e.g., Moreno-Cruz, 2013; Ricke et al. 2013).

3.2 Results for the Default Model Setting Under Empirical Calibration

Under empirical calibration in the default model setting and assuming equal decision-making weights, the optimal SG level among the selected 19 countries is $3.84 Wm^{-2}$. The Nash equilibrium is characterized by a SG level of 8.39 Wm^{-2}, deployed by Venezuela. In the case where decision-making weights are proportional to each country's population, the sample optimal SG level is reduced to 3.16 Wm^{-2}.

Yet the reality is that countries differ in their ability to exert their influence at a global scale—a difference that has been quantified by, e.g., the World Power Index (WPI). We use WPI as weights in our optimization problem and arrive at the following results. First, the sample optimal SG level falls to 3.37 Wm^{-2}, which reflects the preference for lower levels of SG of more powerful countries. Venezuela, Saudi Arabia, Bangladesh, Nigeria, Congo, Brazil, Indonesia, Egypt, and India prefer larger SG levels and thus are potential free drivers. USA, China, and Japan prefer lower SG levels, while the preferences of Australia,

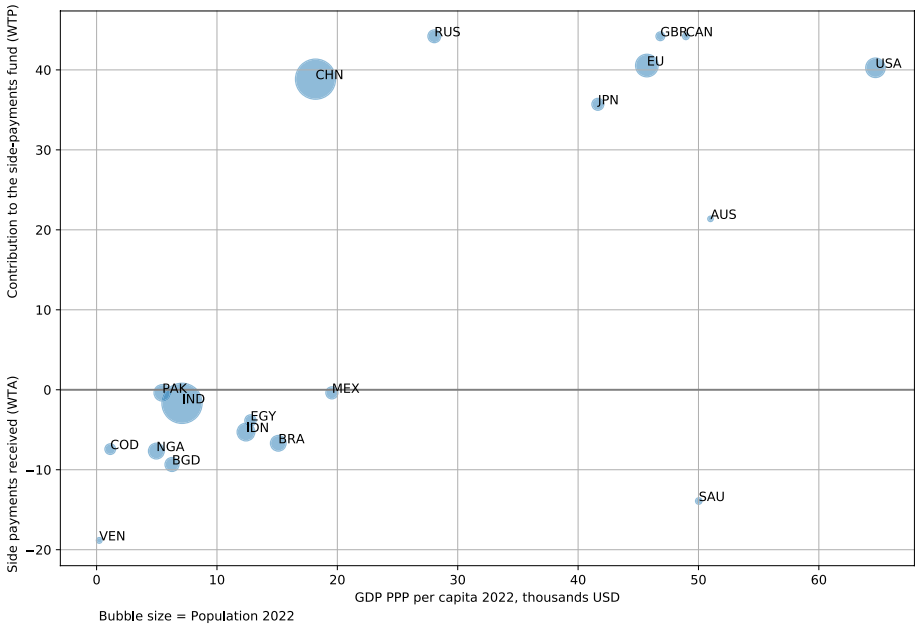

Fig. 2 Contribution to the side-payments fund/side payments received in billions USD (As mentioned in the model parametrization section, these numbers are for illustrative purposes only and the estimates of the actual size of side payments would require a more precise estimates of regional climate damages. While we stress the qualitative nature of our statements, we also note that our estimates of the side payments are conservative, as we underestimate the WTP of non-drivers, which are mostly countries with high GDP per capita, and overestimate the WTA of free drivers, which are mostly countries with low GDP per capita.) (y-axis) and GDP PPP per capita in 2022 in thousands of USD (x-axis) for the 19 countries. The size of the bubble indicates the relative size of the countries' population in 2022

Pakistan, and Mexico are the closest to the optimal level. Russia, Canada, and the UK prefer no SG.

Results of WPI-weighted optimization are illustrated in Fig. 2, which depicts our back-of-the-envelope calculation of contributions to the fund (positive values) estimated as WTP of non-drivers and side payments received (negative values) estimated as WTA of potential free drivers. In reality, a more sensible transfers scheme may be designed that excludes countries with large GDP per capita from receiving side payments (for example, Saudi Arabia) due to their larger capacity to adapt to climate change. For these countries, other instruments of deterrence such as tariffs or sanctions may be used. Similarly, countries with lower GDP per capita may be excused from contributions to the fund or asked for alternative forms of assistance.

4 Results for Alternative Model Specifications

Here we investigate the implications of alternative model specifications for our results. More specifically, we explore modifications to assumptions [A2]–[A4]. We also introduce additional considerations, including the implications of (i) mitigation, (ii) universal fixed costs associated with SG deployment, and (iii) countries' weights in global

decision-making. The game is solved using the same approach as in the default setting. That is, a self-enforcing stable agreement can be reached when the bargaining zone is non-negative. We chose to not overwhelm the Figures that follow with the results for all partial cooperation cases and only show the results for the global coalition.

4.1 Asymmetric Damage Function

Over-provision of SG brings not only additional side effects but could also bring RF-mediated climate impacts that may be worse than climate impacts associated with the under-provision of SG of the same magnitude. More broadly, damage functions may be asymmetric around the optimal SG level. Accordingly, Weitzman (2015) compared the risks of too much SG to a Type-I error, and the loss from too little SG to a Type-II error and proposed an asymmetric piecewise linear loss function. Under the null hypothesis that geoengineering is undesirable, it is reasonable to assume that while the under-provision of SG may be unfortunate since is suboptimal, an overprovision of SG would be considered catastrophic. That is, for each country, the damage function should be asymmetric around its optimal SG level.

We introduce this asymmetry by adding a country-specific quadratic term that takes on a value of zero below a country's optimal SG level, k_i, and increases with the SG level above its optimal value. This new damage function is:

$$D_i^{asym}(G_N) = \frac{\alpha}{2}(G_N - k_i)^2 + \frac{\beta}{2}(max(0, G_N - k_i))^2,$$

where parameter $\beta > 0$ determines the asymmetry.[3] An illustration of this function relative to the default is presented in "Appendix B".

Figure 3 illustrates the implications of asymmetry, with larger values of β being associated with greater asymmetry. For a more intuitive illustration, instead of β in the x-axis, we use the associated increase in the angle of the tangent to the damage function at the point SG = 1, in degrees. The explanation of this measure is presented in the "Appendix B", where you can also find an illustration for an alternative measure of asymmetry, $(\beta - \alpha)/\alpha$.

We see that the socially optimal level of SG decreases with the degree of asymmetry in the damage function; the value in the non-cooperative (Nash) equilibrium does not change (Fig. 3). At the point where damages increase at a prohibitively fast rate beyond the optimal point, a welfare-maximizing strategy is realized by reducing damages in a way that makes no one worse off. This corresponds to the amount of SG that is preferred by the country with the lowest optimal level of SG. In our calibration, these are 2.6, 1, and 1 Wm^{-2} in distributions K^L, K^U, and K^R, respectively and no SG in the empirical distribution K^E. Optimal SG levels converge to these as the degree of asymmetry increases beyond the range presented in Fig. 3.

The greater the damages beyond the optimal point, the higher the stakes in reaching an agreement. Accordingly, the total benefits for non-drivers from cooperation (relative to no cooperation) increase. That is, the maximum amount non-drivers are willing to

[3] Note that in this case the overall damages from SG are also increased. To demonstrate that the results are not crucially dependent on the impact of the change in total damages, we also consider an opposite formulation: reduction of the damages for SG deployment levels below optimal point. This case is presented in "Appendix B" and demonstrates a similar impact of the degree of asymmetry on the ZOPA.

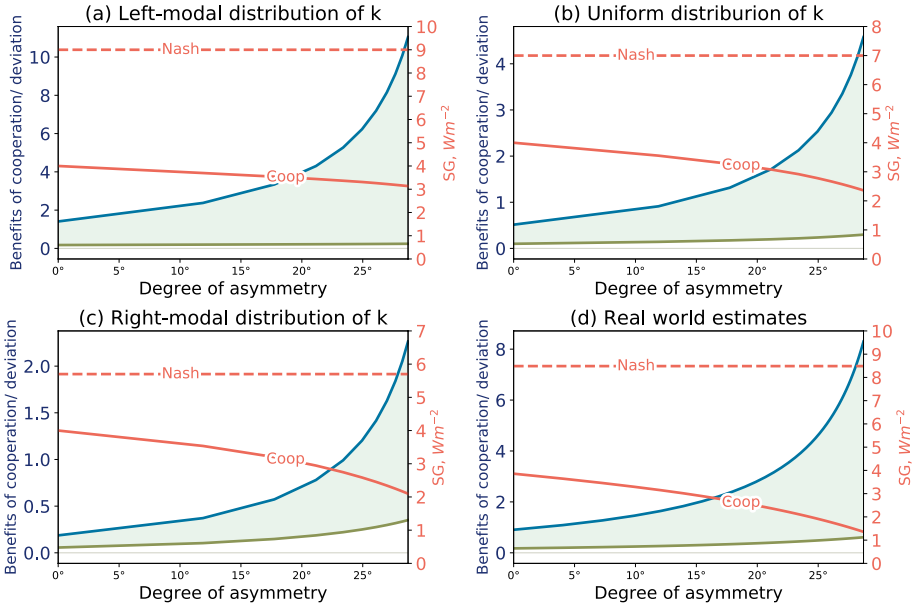

Fig. 3 Left axis: non-drivers' total benefits from cooperation and potential drivers' total benefits from unilateral deviation, both estimated relative to global damages in the no-SG case, are indicated by the blue and the green lines, respectively. The shaded areas between these curves represent the zones of possible agreement. Right axis: the socially optimal levels of SG are shown by the solid red lines and the non-cooperative levels of SG are shown by the dashed red lines. Horizontal axis: the level of damage functions' asymmetry. Four subfigures represent alternative distributions of countries' preferred SG level k_i: **a** left-modal, **b** uniform, **c** right-modal, and **d** empirical

pay to deter free-driving increases with greater asymmetry (Fig. 3, blue line). The lower socially optimal SG level also implies larger benefits to drivers from free driving. That is, the minimum amount drivers are willing to accept to abstain from free driving also increases with greater asymmetry (Fig. 3, green line). While both values increase with the degree of asymmetry, benefits to non-drivers increase more rapidly, reflecting the disproportionately larger negative impact from the over-provision of SG. As a result, the size of the ZOPA (the shaded area in Fig. 3) gets larger with greater asymmetry, improving the chances of reaching an agreement, as even larger side payments are justified to ensure the stability of the global SG agreement. This finding holds for all four distributions of k_i.

4.2 Non-Negligible Deployment Costs

In contrast with current estimates and our default case, here we assume that SG deployment costs are positive and non-negligible. We do this to capture the associated game-theoretic caveats presented by Heyen (2016) and Heyen et al. (2019). In contrast with those works, which model two countries, we consider n heterogeneous countries.

We introduce a deployment cost function that is convex in the individual SG deployment level for each country:

$$C_i(g_i) = \frac{z}{2} g_i^2,$$

where $z > 0$ is the slope of the SG deployment marginal cost function and is the same for all countries. The deployment costs are added to the default climate damage functions (Eq. (1)).

The detailed derivation of the analytical solution is offered in "Appendix C1", and below we summarize our findings and offer an illustration. With non-negligible SG deployment costs, not surprisingly, the socially optimal SG level is lower than in the default setting. Here, as the slope of the marginal cost function is the same for all countries, SG deployment efforts are equally distributed between coalition members ("Appendix C2"). In the non-cooperative equilibrium, the realized SG levels decrease with z (Fig. 4). For all distributions considered here other than the empirical, above some level of direct deployment costs, the non-cooperative equilibrium would result in less SG than the globally optimal levels. Using the empirical calibration, the increase in deployment costs over the range considered, resulted in all countries except for Mexico and Pakistan, which preferences are the closest to the global optimum, abstaining from free driving even without side payments. In this context, free riding refers to the incentive to decrease or cease contribution to the cost of SG deployment. Hence, we refer to potential free riders as *riders.*

Cost-sharing considerations alter the dynamics of the game, balancing the free-driving and free-riding incentives. In the stability analysis, we consider all countries that have the incentive to deviate unilaterally from global cooperation (both *drivers* and *riders*), and we consider the benefits from cooperation to all countries that are better off in the coalition. We refer to countries without an incentive to deviate—neither free ride nor free drive—as *non-deviators.* To reflect the altered incentive structure, the illustration of this case (Fig. 4) is different from the previous ones. We now show benefits to deviators and non-deviators, instead of drivers and non-drivers.

Our simulation results indicate that for any slope of the marginal deployment cost function, z, benefits to non-deviators from cooperation exceed total benefits to deviators from their unilateral deviation. That is, the bargaining zone is non-negative (i.e., ZOPA exists). At lower values of z, ZOPA is maintained by benefits to non-drivers from cooperation. This is similar to the default case, where the challenge is to deter free driving, with the only difference being that the non-cooperative SG level is now lower. Accordingly, side payments can be used to stabilize an agreement.

At larger values of z, ZOPA is maintained by the benefits to potential drivers from cooperation. In other words, countries with high SG preferences benefit from cooperating, with the benefits exceeding those in unilateral deviation or the non-cooperative equilibrium. This is the case because their high SG preferences push the SG level of the coalition up, with deployment costs shared among all members of the deployment coalition. In other words, large deployment costs curb free driving and limit free-riding incentives. The illustration of individual SG deployment levels and corresponding losses of countries in all considered coalition structures is provided in "Appendix C2".

4.3 Counter Solar Geoengineering (CSG)

Heyen et al. (2019) showed that the ability of countries to counteract SG deployment can improve SG coordination. The reason for this is the potential interest of parties to avoid a mutually destructive outcome. This is somewhat similar to the nuclear non-proliferation

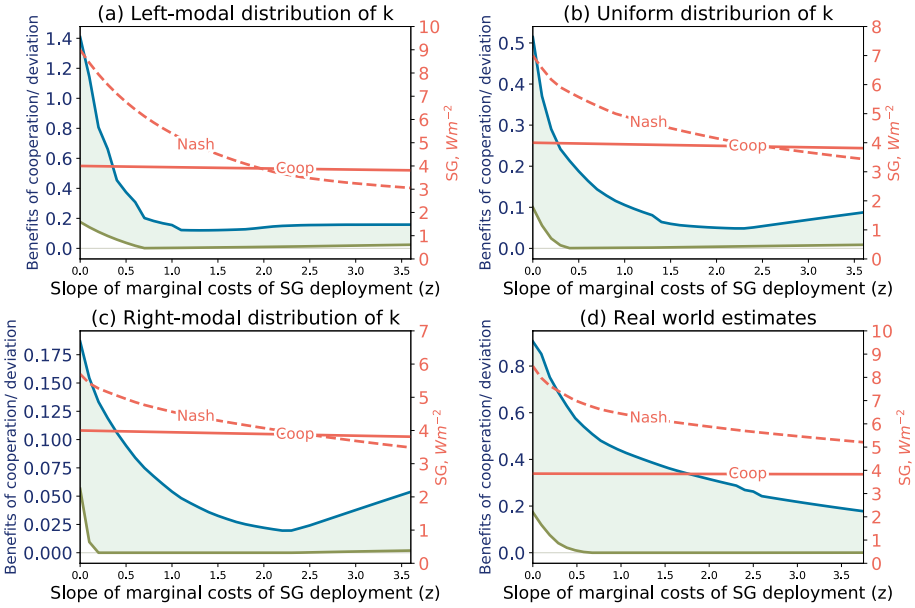

Fig. 4 Left axis: non-deviators' total benefits from cooperation and potential deviators' total benefits from unilateral deviation, both estimated relative to global damages in the no-SG case, are indicated by the blue and the green lines, respectively. The shaded areas between these curves represent the zones of possible agreement. Right axis: the socially optimal levels of SG are shown by the solid red lines and the non-cooperative levels of SG are shown by the dashed red lines. Horizontal axis: the slope of marginal costs of SG deployment. Four subfigures represent alternative distributions of countries' preferred SG level k_i: **a** left-modal, **b** uniform, **c** right-modal, and **d** empirical

regime, where countries understand that the spread of nuclear weapons has negative consequences for both states with and without nuclear weapons, and a treaty represents a common interest in avoiding a nuclear war. Considering counter solar geoengineering (CSG) in the analysis is merited only if the costs of SG deployment are non-negligible. In the absence of such deployment costs, an equilibrium cannot be reached.

Here, in line with Heyen et al. (2019), we introduce CSG as a negative contribution to the amount of SG deployed. We assume that all countries are capable of launching CSG. We use the loss function from the previous specification, but now the range of possible SG values is extended to include negative values, i.e., $g_i \in R$. Fig. 5 illustrates the implications of CSG on the ZOPA, presented under alternative upper limits, L, on CSG: $g_i \geq -\mathrm{L}$. The case $\mathrm{L}=0$ corresponds to the benchmark "non-negligible deployment costs" case with z set at 0.1.

With an option to counter-SG, the cooperative SG level remains the same. Meanwhile, the non-cooperative level decreases since non-drivers can actively oppose SG deployment. The decrease is more rapid with the left-modal distribution where most countries are non-drivers. In all distributions, the non-cooperative SG level is below the socially optimal level when counter-SG is unlimited. The solution for this case and an illustration of the countries' levels of SG deployment and losses are provided in "Appendix D".

Figure 5 shows that larger CSG potential is associated with greater benefits from cooperation. Incentives to deviate disappear at a relatively low limit on CSG. In effect, the

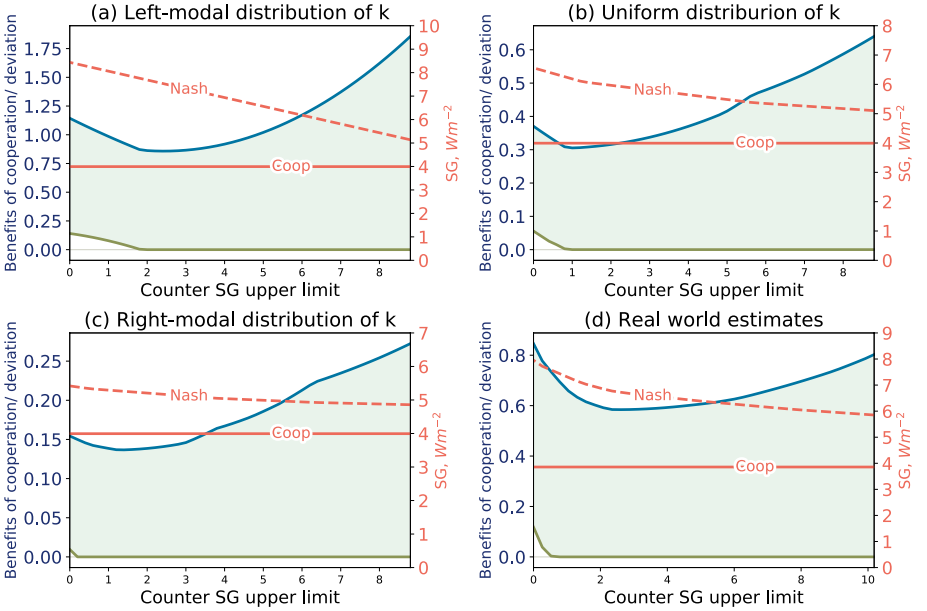

Fig. 5 Left axis: non-deviators' total benefits from cooperation and potential deviators' total benefits from unilateral deviation, both estimated relative to global damages in the no-SG case, are indicated by the blue and the green lines, respectively. The shaded areas between these curves represent the zones of possible agreement. Right axis: the socially optimal levels of SG are shown by the solid red lines and the non-cooperative levels of SG are shown by the dashed red lines. Horizontal axis: the upper limit on counter-SG deployment. Four subfigures represent alternative distributions of countries' preferred SG level k_i: **a** left-modal, **b** uniform, **c** right-modal, and **d** empirical

availability of CSG deters free driving, but when CSG is unlimited in its capacity, it brings a significant waste of resources. This is reflected by the large benefits from cooperation in Fig. 5. In addition, CSG may be hazardous, and therefore not promising.

4.4 Additional Considerations

We next examine the implications of three additional considerations, namely those associated with mitigation, fixed costs of SG deployment, and countries with distinct decision-making weights.

4.4.1 Mitigation and Solar Geoengineering

Mitigation may bring about a reduction in the optimal level of SG for each country. The reduction level $\overline{A_N} \in [0, 1]$ indicates the extent of mitigation, with $A_N = 1$ signifying full decarbonization. Although abatement is exogenous in this setting, we account for the associated costs, which are given[4] as follows: $\frac{c}{2.6}\overline{a}_i^{2.6}$, with the new total cost function being:

[4] Following the dynamic integrated assessment model of climate and economy (DICE), we state the abatement costs increase stronger than linearly with the abatement level.

$$D_i(G_N) = \frac{\alpha}{2}\left(G_N - k_i \bullet \left(1 - \overline{A_N}\right)\right)^2 + \frac{c}{2.6}\overline{a}_i^{2.6}.$$

Derivations are offered in "Appendix E", with results illustrated in Fig. 6. Mitigation reduces the amount of desired SG for each country, and thus both the non-cooperative equilibrium and socially optimal values of SG decrease. As a result, the benefits from cooperation to non-drivers decrease quadratically. The benefits from deviation to non-drivers decline too, albeit at a lower rate. Therefore, the ZOPA narrows with stronger mitigation. This means that stronger mitigation efforts not only reduce climate damage, but also the costs of sustaining a stable and efficient international agreement on SG.

While we do not focus on a specific side-payments scheme, a lower ZOPA may result in lower payments to free drivers, which may also reduce their incentives to decrease emissions abatement. Since abatement is treated as exogenous here, and such considerations are beyond the scope of this study.[5]

4.4.2 Fixed Costs Associated with SG Deployment

There may be fixed costs (FC) of using SG technology. These FC can be interpreted as a consequence of the transition towards a new state with SG in place, with society taking a risk as the transition occurs (Bunn 2019; Wilson 2021).

We assume that FC are non-excludable and the same for all countries, independent of who deploys SG and is given by:

$$FC(G_N) = \begin{cases} 0, G_N = 0; \\ FC, G_N > 0. \end{cases}$$

This yields an augmented damage function:

$$D_i^{FC}(G_N) = \frac{\alpha}{2}(G_N - k_i)^2 + FC(G_N)$$

At a lower end of FC values, the non-cooperative equilibrium and socially optimal SG levels remain unchanged (Fig. 7). While FC brings additional costs, both benefits to non-drivers from cooperation and benefits to drivers from deviation are unchanged. If fixed costs are substantial, a moratorium may be preferred. In the global cooperation case, a moratorium becomes optimal at $C_1 = 0.5\alpha G^{*2}$, while in the non-cooperative case, at a larger value, $FC \geq C_2 \equiv 0.5\alpha k_h^2$. For $FC \geq C_2$, a moratorium emerges as the preferred outcome under a stable, no-side-payments, self-enforcing international agreement, as well as under no regulation.

A particularly challenging situation for an international agreement is where the FC value is in the range: $C_1 \leq FC < C_2$. Here, a moratorium is a socially optimal solution, but drivers prefer more SG. The associated substantial global losses warrant an increased willingness to pay to deter free driving. At the same time, zero socially optimal SG implies larger benefits to drivers from free driving, raising the minimum value drivers are willing to accept to agree to a moratorium. At this point, a non-negative ZOPA still exists. As FC increases closer toward C_2, incentives to free drive decrease, and the amount non-drivers are willing to pay to establish a moratorium increases linearly. Thus, the ZOPA widens.

[5] We leave endogenous mitigation for possible future study.

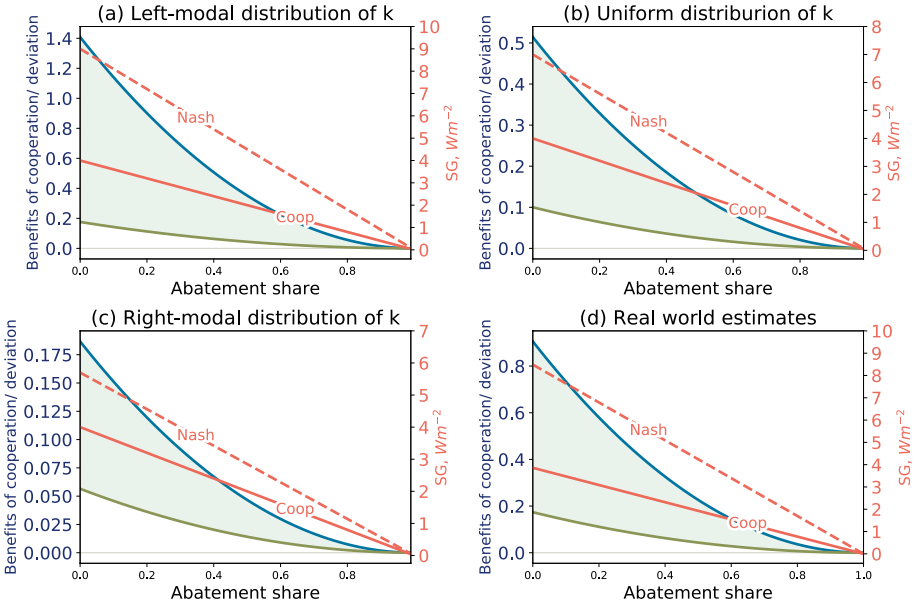

Fig. 6 Left axis: non-drivers' total benefits from cooperation and potential drivers' total benefits from unilateral deviation, both estimated relative to global damages in the no-SG case, are indicated by the blue and the green lines, respectively. The shaded areas between these curves represent the zones of possible agreement. Right axis: the socially optimal levels of SG are shown by the solid red lines and the non-cooperative levels of SG are shown by the dashed red lines. Horizontal axis: the exogenous global mitigation stringency. Four subfigures represent alternative distributions of countries' preferred SG level k_i: **a** left-modal, **b** uniform, **c** right-modal, and **d** empirical

With this, international side payments can be used to stabilize global cooperation regardless of the distribution of preferences. The formal analytical proof is available in "Appendix F".

4.4.3 Countries with Distinct Decision-Making Weights

According to Bunn (2019), a key lesson from the nuclear non-proliferation regime is that a decision-making rule should avoid giving too much weight to a large number of small states. In this spirit, relaxing the assumption that countries have the same weight in decision-making, we introduce country-specific weights, w_i, which are normalized to add to one: $\sum_{i\epsilon N} w_i = 1$. The weight may reflect a country's GDP, population, political capital, or, following Schelling's conflict theory, the capacity to harm other states. In Sect. 3.2 we consider the world power index to find the weighted optimal SG level for our empirical example.

In this setting, the minimization problems of individual countries are unaffected, while the global coalition now minimizes the weighted sum of damage functions:

$$D_N(G_N) = \sum_{i\epsilon N} w_i \left(\frac{\alpha}{2} (G_N - k_i)^2 \right).$$

The result is that the aggregate SG deployment level is a weighted average of countries' optimal SG levels:

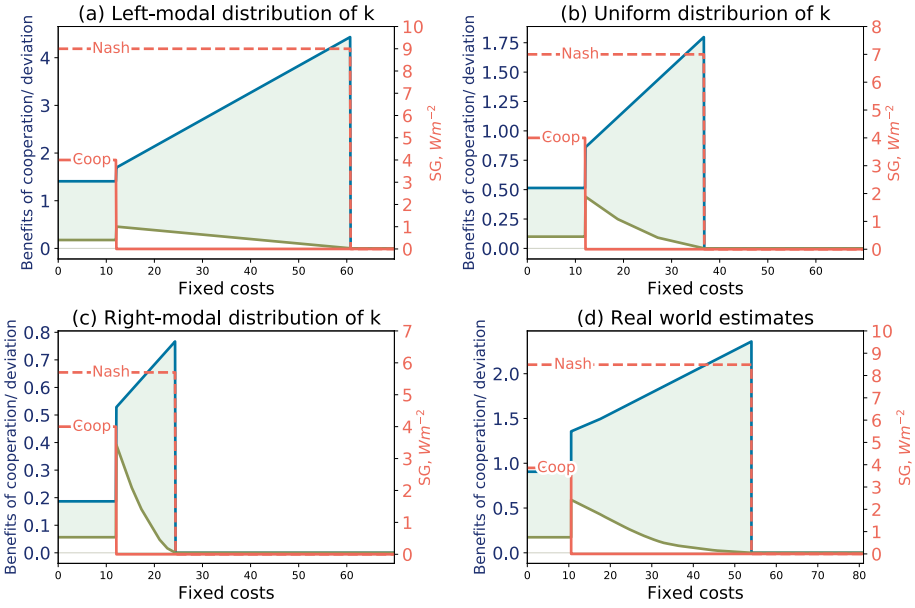

Fig. 7 Left axis: non-drivers' total benefits from cooperation and potential drivers' total benefits from unilateral deviation, both estimated relative to global damages in the no-SG case, are indicated by the blue and the green lines, respectively. The shaded areas between these curves represent the zones of possible agreement. Right axis: the socially optimal levels of SG are shown by the solid red lines and the non-cooperative levels of SG are shown by the dashed red lines. Horizontal axis: fixed costs FC associated with SG deployment. Four subfigures represent alternative distributions of countries' preferred SG level k_i: **a** left-modal, **b** uniform, **c** right-modal, and **d** empirical

$$G^* = \sum\nolimits_{i \in N} w_i k_i.$$

In this case, as in the default setting, an agreement can be stabilized using side payments. The analytical proof is presented in "Appendix G".

5 Conclusions

We have systematically explored the role of international side payments to foster stability of potential efficient international agreements on solar geoengineering. For this, we developed a simple game-theoretic framework, which represents key aspects of SG, and analyzed the possibilities for reaching a stable agreement under alternative model assumptions and parametrizations.

Our results may be said to be optimistic in that our analytical solution indicates that stability of an efficient international SG agreement can be achieved for all distributions of countries' preferred SG levels by means of side payments directed from countries less vulnerable to climate change to those more vulnerable. The effect of such payments is enabled by large asymmetries between countries' vulnerability to climate change and, accordingly, preferred SG level.

Because most studies indicate that vulnerability to climate change is greater for countries with lower GDP (Tol et al. 2004; Mendelsohn et al. 2006), these side payments are

essentially directed from higher- to lower-income countries. This helps to address equity concerns that shape climate negotiations between industrialized countries in the Global North and emerging economies in the Global South.[6] We leave the question of the feasibility of such a side payments scheme to further international relations and political science research but offer our thoughts.

The side payments, where sufficient, would change the cooperation problem from deterring free drivers from over-provision of SG to deterring free riding in contributions to the fund. What gives us a glimmer of hope is progress on an adaptation fund as seen in COP-27. In particular, the established Loss and Damage Fund aims to provide financial assistance to nations most vulnerable to and impacted by the effects of climate change (United Nations Environment Program, 2022). Such a fund is what may be used to deter the most vulnerable countries—potential free drivers—from SG over-provision.

A moratorium on SG is commonly discussed in the context of SG governance, and our results indicate that a moratorium is an outcome of an efficient agreement only when SG deployment is associated with non-excludable fixed costs (FCs). When FCs are large enough to deter free driving, side payments are not needed for a stable agreement on the moratorium. In other cases, a stable SG moratorium agreement is attainable by means of side payments.

While we do not consider interactions between SG and mitigation, we look at the implications of the level of emissions abatement on SG agreement and find that greater mitigation reduces the cost of sustaining a stable agreement because it reduces both the maximum amount drivers are willing to pay to deter free driving and the minimum amount drivers are willing to accept to abstain from free driving.

The "optimistic results" we present should be taken with a grain of salt. First, the choice of a SG strategy is more complicated than we have characterized it, as it would involve more than the quantity of SG, but also the timing, frequency, geographic location, and other details, e.g., materials used, of deployment. Current scientific evidence on regional SG impacts is fragmented at best, especially when it comes to alternative specifications of regionally focused SG deployment strategies. Second, our model is static and does not capture the dynamic structure of coalition formation (Heyen and Lehtomaa 2021); nor does it capture political realities that may go well beyond simple economic considerations. Third, SG impacts are subject to deep uncertainty, which need to be accounted for in the design of any side payments scheme. As an example, looking at SG decision-making under deep uncertainty about SG impacts, Manoussi et al. (2018) find that the level of SG deployed decreases in both cooperative and Nash equilibria with greater uncertainty. Fourth, we treat countries as acting as rational economic agents, which seek to minimize climate damage. In a world where countries may pursue other geopolitical interests, positive economic incentives—such as side payments—may not be sufficient to guarantee the absence of unilateral deployment of SG above the socially optimal level. Fifth, we consider SG in isolation from mitigation. Linking a SG agreement with mitigation may address one of the most prominent concerns about SG, namely the moral hazard that mere consideration of SG as a policy option would reduce incentives for emissions cuts. Finally, we consider only unilateral deviations from global cooperation. In the case of non-negligible deployment costs,

[6] For comparison, side payments that enhance stability in international mitigation agreements suggest the opposite direction of side payments (Carraro et al. 2006; McGinty 2006), which is hardly justifiable from equity considerations and historical emissions responsibility.

an interesting extension of the analysis would be to consider deviations by more than one country with the subsequent formation of new coalitions to share the deployment costs.

Despite the abovementioned limitations, we would like to emphasize the key role that adaptation assistance to countries most vulnerable to climate change could play in facilitating the design of a stable international SG agreement regime.

A Default Specification

In the default model specification, damage function for country $i \in N, i = 1, 2, \ldots n$, reads:

$$D_i = \frac{\alpha}{2}(G_N - k_i)^2,$$

where G_N is realized SG level and k_i - country $i's$ optimal SG level. The damage function can be rewritten as:

$$D_i = \frac{\alpha}{2}G_N^2 - \alpha G_N k_i + d_i$$

with $d_i \equiv \frac{1}{2}\alpha k_i^2$ - country $i's$ climate damages in the absence of SG.

In the following analytical derivations we offer a solution in a more general form, where parameter d_i is explicit. It must, however, satisfy the following condition:

$$d_i \geq \frac{1}{2}\alpha k_i^2.$$

This condition ensures that damages cannot be negative, i.e., SG cannot be used to bring benefits beyond reduced climate damages. The value d_i such that $d_i > \frac{1}{2}\alpha k_i^2$ describes the case when the damage compensation potential of SG differs across countries. For example, $d_i = \frac{1}{2 \cdot 0.9}\alpha k_i^2$ means that SG can be used to compensate for 90% of climate damages in each country (in each - different level of SG). Here, residual damages would be largest (in terms of its levels) for the country most vulnerable as illustrated in Fig. 8.

A.1 Third-Stage Solution

Signatories

Consider an arbitrary coalition $S \subseteq N$. Total SG deployment of coalition members amounts to $G_S = \sum_{i \in S} g_i$; the average level of SG preferences of coalition members is $\bar{k}_S = \sum_{i \in S} k_i / s$, where s - number of coalition members.

In the third stage of the game, coalition members minimize their total damages D_S:

$$\min_{G_S} D_S = \frac{1}{2}s\alpha(G_S + G_{N \setminus S})^2 - \alpha(G_S + G_{N \setminus S}) \sum_{i \in S} k_i + \sum_{i \in S} d_i$$

Coalition S reaction function:

$$G_S = \begin{cases} \bar{k}_S - G_{N \setminus S} & if\ G_{N \setminus S} < \bar{k}_S \\ 0 & if\ G_{N \setminus S} \geq \bar{k}_S \end{cases}$$

Non-signatories

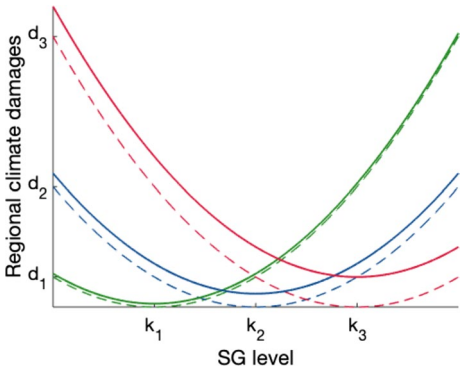

Fig. 8 Larger d_i. Default damage functions (dashed line) versus damage functions that assume that SG can be used to reduce 90% of region $i's$ climate damages when SG is used at a level that is optimal for the region i (solid line)

The non-signatory $i \in N \setminus S$ minimizes its individual damage function:

$$\min_{g_i} D_i = \frac{1}{2}\alpha(g_i + G_{N\setminus i})^2 - \alpha k_i(g_i + G_{N\setminus i}) + d_i$$

The non-signatory i's reaction function reads:

$$g_i = \begin{cases} k_i - G_{N\setminus i} & if\ G_{N\setminus i} < k_i \\ 0 & if\ G_{N\setminus i} \geq k_i \end{cases}$$

Total SG Deployment Level and Associated Damage Costs

Consider country j to be the country with the largest preferred SG level among non-signatories. Thereby, an arbitrary coalition S may be"active", i.e. define the global thermostat, if it's optimal deployment level exceeds k_j and "passive" otherwise.

	Active coalitions:$\bar{k}_S > k_j$	Passive coalitions:$\bar{k}_S < k_j$
G_N	$\bar{k}_S$	k_j
D_i	$\alpha\bar{k}_S(0.5\bar{k}_S - k_i) + d_i$	$\alpha k_j(0.5k_j - k_i) + d_i$
D_N	$\alpha n\bar{k}_S(0.5\bar{k}_S - \bar{k}_N) + d_N$	$\alpha nk_j(0.5k_j - \bar{k}_N) + d_N$

A.2 The Proposition Proof

Consider an arbitrary active coalition $S \subseteq N$. Optimal level of SG for this coalition is $\bar{k}_S$. Let us denote by DR^S the set of drivers: $j \in DR^S$ if $j \in S$ and $k_j > \bar{k}_S$, and by ND^S the set of non-drivers: $i \in ND^S$ if $i \in S$ and $k_i \leq \bar{k}_S$.

The proposition states that the total gain of non-drivers from cooperation (relative to the no-cooperation case) exceeds the sum of the gains of all drivers from their unilateral deviation from the cooperation:

$$\sum_{i \in ND^S} (D_i(\emptyset) - D_i(S)) \geq \sum_{j \in DR^S} (D_j(S) - D_j(S \setminus j))$$

Substitute the considered functional form of damages, here k_h denotes the highest SG preference of all considered countries from set N:

$$\frac{\alpha}{2}\sum_{i\in ND^S}\left(k_h-\bar{k_S}\right)\left(\bar{k_S}+k_h-2k_i\right)\geq\frac{\alpha}{2}\sum_{j\in DR^S}\left(k_j-\bar{k_S}\right)^2$$

Proof By definition, k_h is the largest value of all nationally-optimal SG levels. It is larger than the average of optimal levels for all non-drivers: $k_h \geq \bar{k}_{ND^S}$. Multiplying both sides by the number of non-drivers in a coalition S, we arrive at:

$$|ND^S|k_h \geq k_{ND^S}$$

where $|ND^S|$ is a number of non-drivers in a coalition S, $k_{ND^S}=\sum_{i\in ND^S}k_i$ is the sum all SG amounts preferred by non-drivers. Then we add them up and subtract from the right hand side (RHS) of the sum of all preferences of coalition members (k_S):

$$|ND^S|k_h \geq k_{ND^S}+\underbrace{(k_{ND^S}+k_{DR^S})}_{k_S}-\underbrace{(|DR^S|+|ND^S|)\bar{k}_S}_{k_S}$$

After some reshuffling we arrive at:

$$|ND^S|k_h+|ND^S|\bar{k}_S-2k_{ND^S}\geq k_{DR^S}-|DR^S|\bar{k}_S$$

We then rewrite this inequality using summation operator as:

$$\sum_{i\in ND^S}\left(k_h+\bar{k}_S-2k_i\right)\geq\sum_{j\in DR^S}\left(k_j-\bar{k}_S\right)$$

.

Multiplying both sides by $\left(k_h-\bar{k}_S\right)\geq 0$, the difference between the largest preferred SG level among all countries in N and the average of coalition member's preferred SG level.

$$\left(k_h-\bar{k}_S\right)\sum_{i\in ND^S}\left(\bar{k}_S+k_h-2k_i\right)\geq\left(k_h-\bar{k}_S\right)\sum_{j\in DR^S}\left(k_j-\bar{k}_S\right)$$

By definition $k_h\geq k_i\ \forall i\in N$, the RHS is greater than or equal to $\sum_{j\in DR^S}\left(k_j-\bar{k}_S\right)^2$. Therefore, the following holds:

$$\left(k_h-\bar{k}_S\right)\sum_{i\in ND^S}\left(\bar{k}_S+k_h-2k_i\right)\geq\sum_{j\in DR^S}\left(k_j-\bar{k}_S\right)^2$$

Finally, by multiplying both sides by 0.5α we arrive at the following expression, which was to be demonstrated:

$$\frac{\alpha}{2}\sum_{i\in ND^S}\left(k_h-\bar{k}_S\right)\left(\bar{k}_S+k_h-2k_i\right)\geq\frac{\alpha}{2}\sum_{j\in DR^S}\left(k_j-\bar{k}_S\right)^2$$

. □

A.3 Illustration

Figure 9 illustrates our Proposition, which itself is in the spirit of Coase theorem. What differs from the Coase theorem case, is that we are dealing with public "gob" and thus

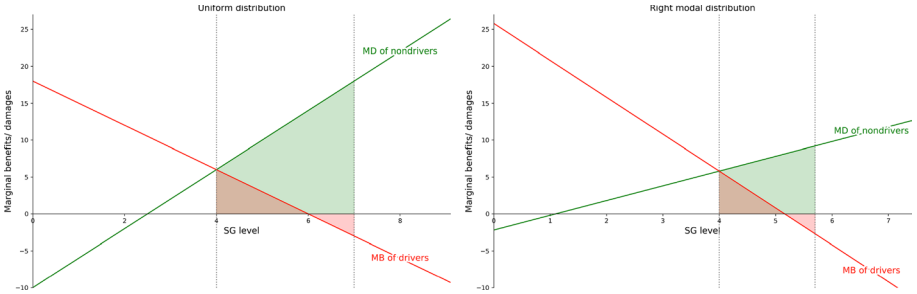

Fig. 9 Total marginal damages to nondrivers (green line) and the sum of marginal benefits to potential free drivers (red line) under a uniform distribution of optimal values $k^U = \{1, 2, 3, 4, 5, 6, 7\}$ (left) and under right-modal distribution $k^R = \{1.0, 1.2, 4.5, 4.8, 5.3, 5.5, 5.7\}$ (right)

country may alter between being a loser or winner dependent on the amount of "gob" supplied.

We consider two groups of countries: (i) non-drivers and (ii) potential free drivers. Note that marginal damages to non-drivers are negative at low SG values, indicating that non-drivers benefit from small amounts of SG due to the associated reduced climate damages. Similarly, marginal benefits to potential free driversis negative when SG levels are large. Thereby, there is a difference between the sum of damages to free drivers in Nash equilibrium (where some free drivers actually lose) and the sum of damages that potential free drivers bear from the unilateral deviations.

In Fig. 9, the green area shows the benefits to non-drivers from global cooperation relative to the Nash equilibrium case. Red area shows the benefits to potential free drivers in Nash equilibrium relative to the global cooperation case. Here, red area (triangle) in the negative zone indicates losses to potential drivers from SG over-provision in Nash equilibrium.

B Asymmetric Damage Function

We define the degree of asymmetry B as a deviation (in degrees) of D_i^{asym} from the default function D_i at a point $G_N = k_i + 1$. The resulting degree of asymmetry is:

$$B(\beta) = (arctan(\alpha + \beta) - arctan(\alpha)) \cdot \frac{180}{\pi} [degrees]$$

where $\alpha = 1.5$ and associated $arctan(\alpha) = 56.3$ [degrees].

We consider two types of asymmetry: with relatively larger damages from SG over-provision and smaller damages from SG under-provision. This helps us to demonstrate the potential for the qualitative change in the results. The illustration of the considered asymmetric damage functions are presented in the Fig. 10.

Changes in the total amount of SG, benefits to ND of cooperation and benefits to DR from unilateral deviation for the left-side asymmetry are depicted in the Fig. 11. This figure demonstrates that ZOPA is positive and increasing with the degree of asymmetry.

We also offer results for the alternative measure of asymmetry, which is as follows: $(\beta - \alpha)/\alpha$. The corresponding graphs are presented in the Fig. 12.

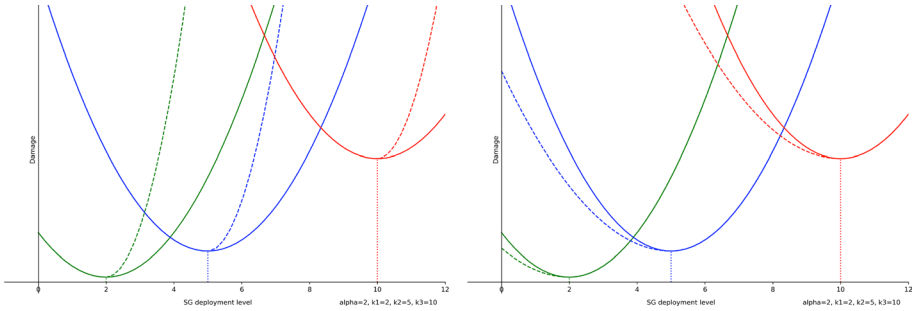

Fig. 10 **a** Asymmetric damage functions with stronger damage for SG overprovision (dashed line) versus default damage functions (solid line). **b** Asymmetric damage functions with lower damage for SG underprovision (dashed line) versus default damage functions (solidline)

C Non-negligible Direct SG Deployment Costs

When assuming that the deployment costs are non-negligible, the damage function reads:

$$D_i = \frac{1}{2}\alpha G_N^2 - \alpha k_i G_N + d_i + \frac{1}{2} z g_i^2$$

where z is the slope of marginal SG deployment costs and $g_i \geq 0$.

C.1 Third-Stage Solution

Signatories

Minimization problem for members of a coalition $S \subseteq N$:

$$\min_{g_i} D_S = 0.5 s\alpha (G_S + G_{N\setminus S})^2 - \alpha k_S (G_S + G_{N\setminus S}) + d_S + 0.5 z \sum_{i=1}^{s} g_i^2$$

FOC:

$$\frac{\partial D_S}{\partial g_i} = s\alpha (G_S + G_{N\setminus S}) - \alpha k_S + z g_i = 0$$

Adding it up for the coalition members $i \in S$, we arrive at:

$$s^2 \alpha (G_S + G_{N\setminus S}) - \alpha s k_S + z G_S = 0$$

Reaction function for a coalition S then reads:

$$G_S = \begin{cases} \frac{\alpha s^2}{\alpha s^2 + z} (\bar{k}_S - G_{N\setminus S}) & \text{if } G_{N\setminus S} < \bar{k}_S \\ 0 & \text{if } G_{N\setminus S} \geq \bar{k}_S \end{cases}$$

Non-signatories

Non-signatory $i \in N \setminus S$ minimizes it's individual damage function as follows:

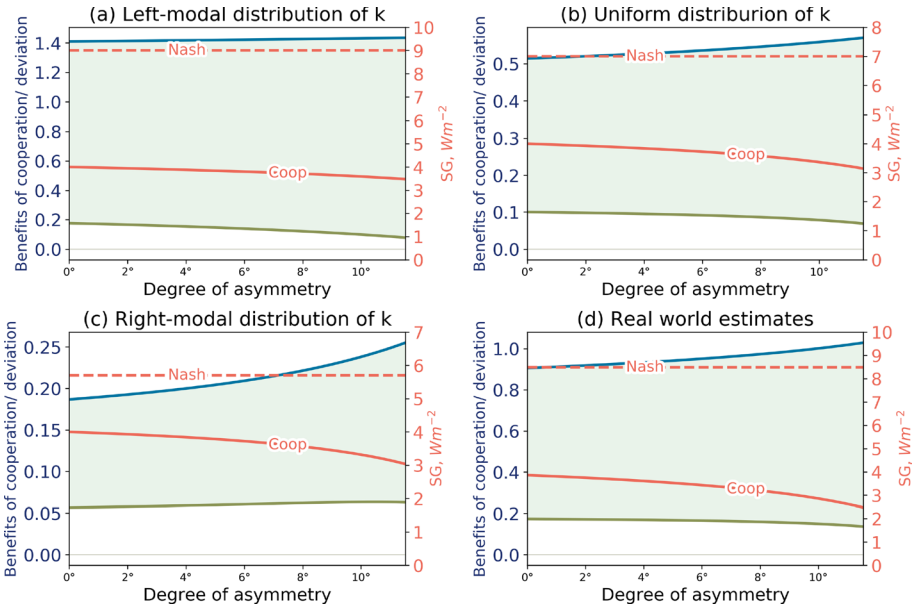

Fig. 11 Changes in socially optimal SG level (solid red line) and non-cooperative SG level (dashed red line) as a function of damage functions' 'left-side' asymmetry, under alternative distributions of countries' preferred SG level k_i, including: **a** left-modal, **b** uniform, **c** right-modal distributions, and **d** empirical distribution. Associated benefits to non-drivers from cooperation (blue line) and total benefits to drivers from unilateral deviation (green line), estimated as a reduction in damages relative to the no-SG case. Shaded area between the curves represents the zone of possible agreement

$$\min_{g_i} D_i = 0.5\alpha(g_i + G_{N\setminus i})^2 - \alpha k_i(g_i + G_{N\setminus i}) + d_i + 0.5zg_i^2$$

FOC:

$$\frac{\partial D_i}{\partial g_i} = \alpha(g_i + G_{N\setminus i}) - \alpha k_i + zg_i = 0$$

The associated reaction function reads:

$$g_i = \begin{cases} \frac{\alpha}{\alpha+z}(k_i - G_{N\setminus i}) & G_{N\setminus i} \le k_i \\ 0 & G_{N\setminus i} > k_i \end{cases}$$

C.2 Selected Illustrations

Figure 13 shows individual and total levels of SG deployment in global cooperation, Nash equilibrium and under the unilateral deviation case. The Figure is presented for a uniform distribution of optimal values $k^U = \{1, 2, 3, 4, 5, 6, 7\}$ for countries 2–7.

For countries with very low preferred SG levels (1 and 2), the deviation always results in zero individual SG deployment and an increase in total deployment, since otherl coalition members have larger average over preferred SG levels. The difference

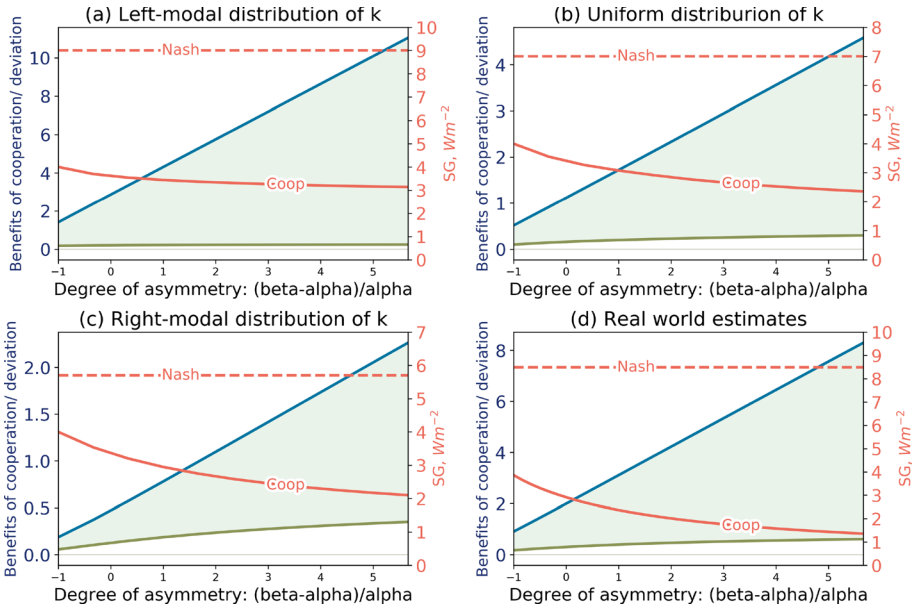

Fig. 12 Changes in socially optimal SG level (solid red line) and non-cooperative SG level (dashed red line) as a function of damage functions' 'right-side' asymmetry measured as $(\beta - \alpha)/\alpha$, under alternative distributions of countries' preferred SG level k_i, including: **a** left-modal, **b** uniform, **c** right-modal distributions, and **d** empirical distribution. Associated benefits to non-drivers from cooperation (blue line) and total benefits to drivers from unilateral deviation (green line), estimated as a reduction in damages relative to the no-SG case. Shaded area between the curves represents the zone of possible agreement

between the total SG deployment in cooperation and deviation decreases with marginal deployment costs, z.

Preferences of countries 3 and 4 are the closest to the socially optimal SG deployment level. Therefore, the difference between total SG level in the cooperative case and their unilateral deviation is tiny. This creates incentives to free ride: individual SG deployment level (deviation) is either zero or is very low.

Countries 5–7 have large preferred SG level and free drive when marginal deployment costs are relatively low: their individual SG deployment level in the case of deviation determines the total SG level. However, increasing costs of deployment curb their free driving ability and the SG deployment level decreases below it's cooperative level. This is the case because the optimal SG level of other coalition members is lower. As a result, these countries do not have an incentive to free drive when SG deployment costs are substantial. However, as deployment costs are increasing, countries may become free riders. In the considered range of marginal deployment costs this happens only for the country 5 shown in Fig. 14.

Figure 14 depicts individual losses in cooperation, Nash equilibrium and unilateral deviation cases for countries 2–7. Unilateral deviation occurs if the associated losses are smaller than in the cooperative case. This difference is depicted in Fig. 14 by the area. The area in green indicates the case when a country is better off in the cooperative state. Figure 14 demonstrates that countries with large preferred SG levels (countries 5–7) have an incentive to deviate when SG deployment costs are low. They also may have an incentive

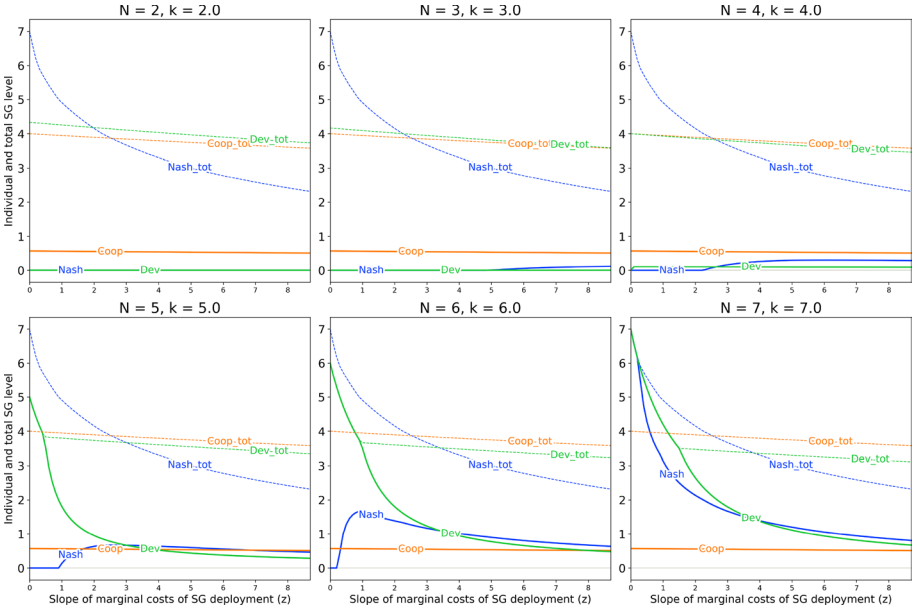

Fig. 13 Effect of the marginal deployment costs on individual and total levels of SG deployment in (i) global cooperation (orange lines); (ii) Nash equilibrium (blue lines); (iii) unilateral deviation (green lines)

to act as free riders if deployment costs are substantial. In the considered range of marginal deployment costs, it happens only for the country 5.

Countries whose preferred SG level is close to the socially optimal value (i.e., countries 3 and especially 4) have the strongest free-riding incentive. For these countries deviation represents costs with no benefits and thus is not rational. Countries with small preferred SG level, such as country 2, have small incentive to free ride at a relatively large deployment costs.

D Counter SG with Non-negligible Deployment Costs

Damage function:

$$D_i = \frac{1}{2}\alpha G_N^2 - \alpha k_i G_N + d_i + \frac{1}{2} z g_i^2$$

Where z is the slope of marginal SG or counter-SG deployment costs. The main difference with the previous case is that g_i may be both positive and negative: $g_i \in R$.

D.1 Third-Stage Solution

Minimization problems are similar to the previous case with non-negligible costs. The difference is in the reaction functions for the case where g_i is not restricted to be non-negative:

Reaction function for a coalition S reads:

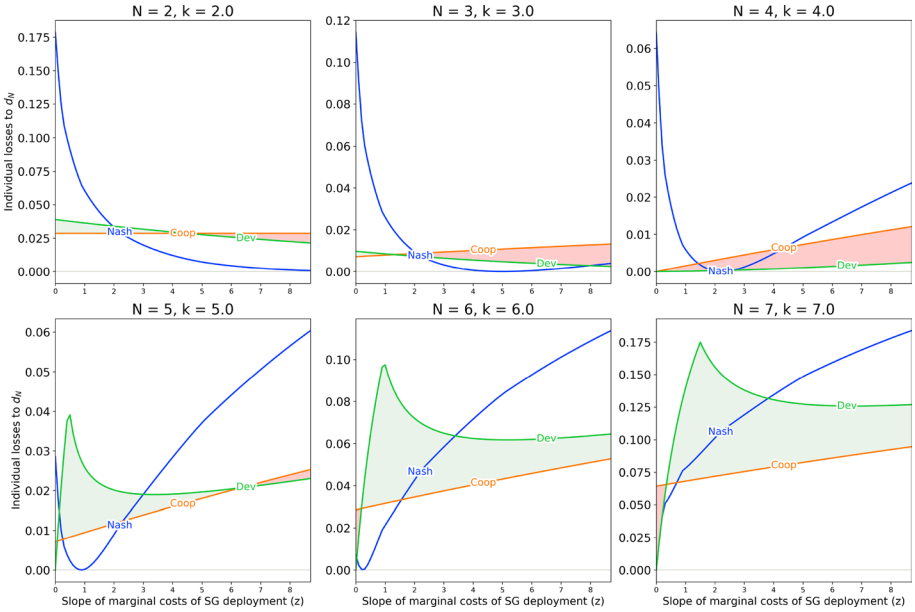

Fig. 14 Effect of the marginal deployment costs on individual losses in (i) global cooperation (orange line); (ii) Nash equilibrium (blue line); (iii) unilateral deviation (green line). Individual incentives to deviate are indicated by the area in red, incentives to cooperate - the area in green

$$G_S = \frac{\alpha s^2}{\alpha s^2 + z}(\bar{k}_S - G_{N\setminus S})$$

Non-signatory $i \in N \setminus S$ reaction function reads:

$$g_i = \frac{\alpha}{\alpha + z}(k_i - G_{N\setminus i})$$

D.2 Illustrations

Figure 15 shows individual SG deployment levels for the scenarios of global cooperation, Nash equilibrium and unilateral deviation, where marginal costs of both SG and counter-SG deployment aret set to $z = 0.1$. The behavior of countries is intuitive: non-drivers counteract SG deployment activities of drivers (subject to a specified limit). In Nash equilibrium (upper graphs) some countries may switch from counter-SG to SG deployment as the CSG limit increases. In unilateral deviation (lower graphs) countries are clearly divided into two groups: (i) drivers that deploy SG beyond cooperative optimum, and (ii) non-drivers that counteract SG deployment. Note that a country whose optimal SG level almost coincides with the global optimal level (country 4 in the uniform distribution, middle column) tends to free ride in the case of a unilateral deviation, avoiding either activity.

Figure 16 depicts countries' individual losses in Nash equilibrium and in the case of a unilateral deviation. For comparison, individual losses in cooperation are indicated by dashed lines. Color stays the same for each country.

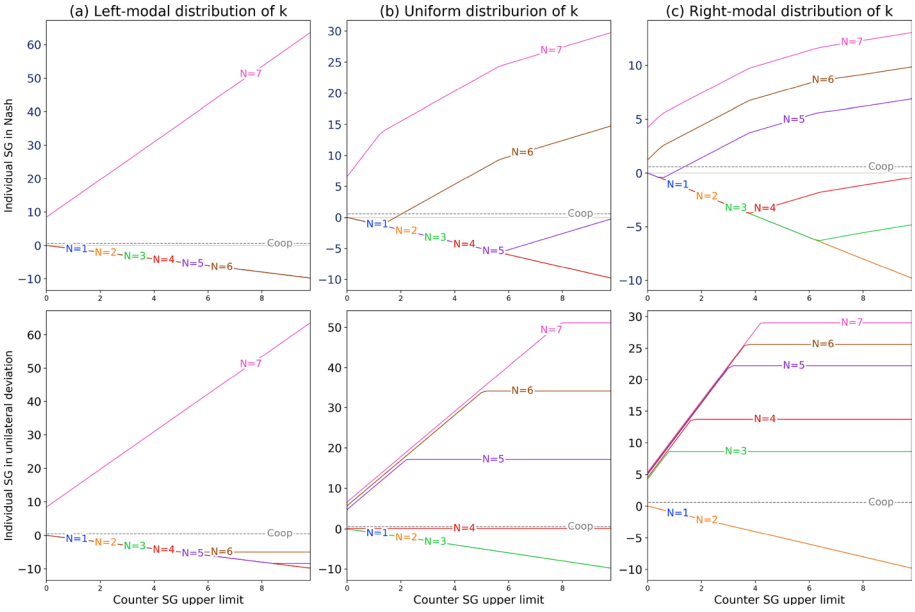

Fig. 15 Impact of the upper limit of counter- SG on individual levels of SG deployment in (i) global cooperation (grey dashed lines); (ii) Nash equilibrium (three upper graphs); (iii) unilateral deviation (three lower graphs). Columns refer to three considered distributions of k: **a** left-modal, **b** uniform, and **c** right-modal distributions

E Exogenous Mitigation

Consider the global share of emissions abatement $\sum_{i \in N} \bar{a} = \bar{A}_N \in [0, 1]$.

Damage function then reads:

$$D_i = \frac{\alpha}{2} G_N^2 - (1 - \bar{A}_N)\alpha k_i G_N + (1 - \bar{A}_N)^2 d_i + \frac{c_i}{2.6} \bar{a}_i^{2.6}$$

Where c_i in the parameter of abatement costs of country i.

Signatories

Minimization problem for members of a coalition $S \subseteq N$ reads:

$$\min D_S = \frac{1}{2} s\alpha (G_S + G_{N \setminus S})^2 - (1 - \bar{A}_N)\alpha k_S (G_S + G_{N \setminus S}) + (1 - \bar{A}_N)^2 d_S + \sum_{i \in S} \frac{c_i}{2.6} \bar{a}_i^{2.6}$$

Non-signatories

Non-signatory $i \in N \setminus S$ minimize their individual damage function as following:

$$\min D_i = \frac{\alpha}{2} G_N^2 - (1 - \bar{A}_N)\alpha k_i G_N + (1 - \bar{A}_N)^2 d_i + \frac{c_i}{2.6} \bar{a}_i^{2.6}$$

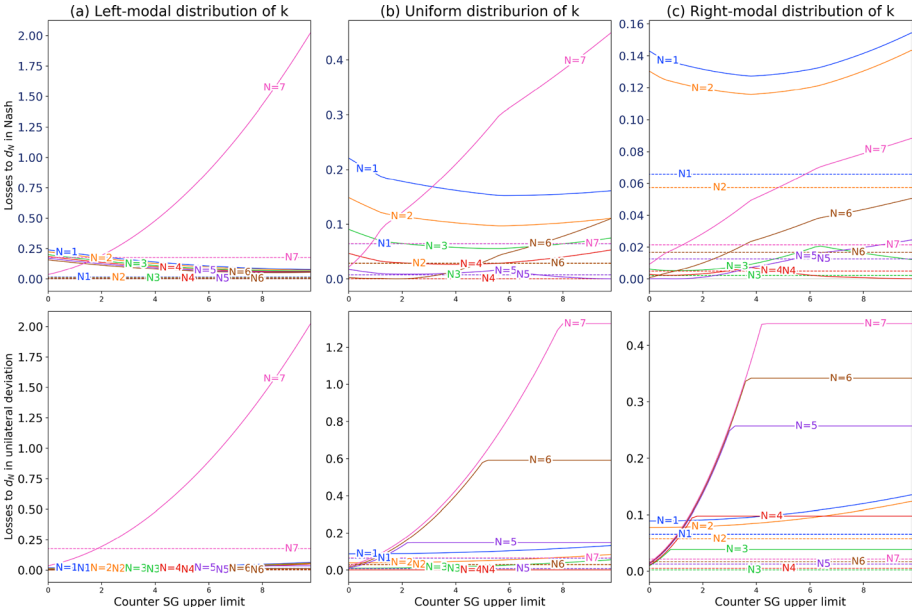

Fig. 16 Impact of the upper limit of counter- SG on individual losses in (i) global cooperation (dashed lines); (ii) Nash equilibrium (three upper graphs); (iii) unilateral deviation (three lower graphs). Columns refer to three considered distributions of k: **a** left-modal, **b** uniform, and **c** right-modal distributions

E.1 Amount of Transfers

Minimum amount of transfers to sustain the cooperation is the sum of benefits from deviation of all free drivers:

$$\sum_{j \in DR} (D_j(N) - D_j(N \setminus j)) = 0.5\alpha(1 - \bar{A}_N)^2 \sum_{j \in DR} \left(k_j - \frac{k_N}{n}\right)^2$$

Note that the actual transfers size may exceed this level.

F Fixed Costs Associated with SG Deployment

Damage function:

$$D_i = \frac{1}{2}\alpha G_N^2 - \alpha k_i G_N + d_i + FC(G_N)$$

Where

$$FC(G_N) = \begin{cases} 0 & if\ G_N = 0 \\ FC & if\ G_N > 0 \end{cases}$$

F.1 The Proposition Proof

The difference between the default specification and the modification with fixed costs of deployment is that a corner solution may come up. Here, corner solution means moratorium on SG. We consider an active coalition S with the optimal deployment level $\bar{k_S}$. We distinguish three cases depending on the magnitude of fixed costs, FC:

- $FC < 0.5\alpha(\bar{k_S})^2$ - Deployment equilibrium (interior solution) in both coalition and no cooperation. The case is similar to the default specification, without FC.
- $0.5\alpha(\bar{k_S})^2 \leq FC < 0.5\alpha k_h^2$ - Moratorium on SG in coalition and deployment in non-cooperative scenario.
- $FC \geq 0.5\alpha k_h^2$ - Moratorium on SG in both coalition and non-cooperative scenario: no collective action problem.

The interesting case (and thus our focus in the following) is when $0.5\alpha(\bar{k_S})^2 \leq FC < 0.5\alpha k_j^2$.

Individual damages	
	$ND: k_i \leq (2FC/\alpha)^{0.5}$
$D_i(S)$	d_i
$D_i(S \setminus j)$	$0.5\alpha k_j^2 - \alpha k_i k_j + FC + d_i$
$D_i(S \setminus j) - D_i(S)$	$0.5\alpha k_j(k_j - 2k_i) + FC$
	$DR: k_j > (2FC/\alpha)^{0.5}$
$D_j(S)$	d_j
$D_j(S \setminus j)$	$-0.5\alpha k_j^2 + FC + d_j$
$D_j(S) - D_j(S \setminus j)$	$0.5\alpha k_j^2 - FC$

Proposition reads: $\sum_{j \in DR^S}(D_j(S) - D_j(S \setminus j)) \leq \sum_{i \in ND^S}(D_i(\emptyset) - D_i(S))$

After we substitute the considered functional form of damages, where k_h denotes the largest preferred SG level of all considered countries from set N, it reads:

$$\sum_{j \in DR^S}(\frac{\alpha}{2}k_j^2 - FC) \leq \sum_{i \in ND^S}(\frac{\alpha}{2}k_h(k_h - 2k_i) + FC)$$

Proof $|DR^S|$ is the number of drivers in a coalition S and $|ND^S|$ is the number of non-drivers in a coalition S.

$$\frac{\alpha}{2}\sum_{j \in DR^S} k_j^2 - |DR^S| \cdot FC \leq \frac{\alpha}{2}|ND^S|k_h^2 - \alpha k_h k_{ND^S} + |ND^S| \cdot FC$$

We can rewrite as follows:

$$\frac{\alpha}{2}\sum_{j \in DR^S} k_j^2 \leq \frac{\alpha}{2}|ND^S|k_h^2 - \alpha k_h k_{ND^S} + sFC$$

Assumming $0.5\alpha(\bar{k_S})^2 \leq FC$, it follows:

$$\frac{\alpha}{2}\sum_{j\in DR^S} k_j^2 \le \frac{\alpha}{2}|ND^S|k_h^2 - \alpha k_h k_{ND^S} + \frac{\alpha}{2}s\left(\bar{k_S}\right)^2$$

Multiply by $2/\alpha$ and rewrite LHS in an identical form:

$$\sum_{j\in DR^S}\left(k_h k_j - (k_h - k_j)k_j\right) \le |ND^S|k_h^2 - 2k_h k_{ND^S} + s\left(\bar{k_S}\right)^2$$

Open the summation operator:

$$k_h k_{DR^S} - \sum_{j\in DR^S}\left((k_h - k_j)k_j\right) \le |ND^S|k_h^2 - 2k_h k_{ND^S} + s\left(\bar{k_S}\right)^2$$

Can be rewritten as follows:

$$-\sum_{j\in DR^S}\left((k_h - k_j)k_j\right) + k_h k_{ND^S} \le sk_h^2 - |DR^S|k_h^2 - k_h k_S + \frac{k_S^2}{s}$$

Combine RHS in the square of a difference:

$$|DR|k_h^2 - \sum_{j\in DR^S}\left((k_h - k_j)k_j\right) + k_h k_{ND^S} - k_h k_S \le s\left(k_h^2 - 2k_h\frac{k_S}{s} + \left(\frac{k_S}{s}\right)^2\right)$$

Rewrite LHS in an identical form:

$$\sum_{j\in DR^S}(k_h(k_h - k_j)) - \sum_{j\in DR^S}\left((k_h - k_j)k_j\right) + k_h k_{DR^S} + k_h k_{ND^S} - k_h k_S \le s\left(k_h - \frac{k_S}{s}\right)^2$$

Cancel out three terms from LHS:

$$\sum_{j\in DR^S}(k_h(k_h - k_j)) - \sum_{j\in DR^S}\left((k_h - k_j)k_j\right) \le s\left(k_h - \frac{k_S}{s}\right)^2$$

The following is true by the definition:

$$\sum_{j\in DR^S}(k_h - k_j)^2 \le s\left(k_h - \bar{k_S}\right)^2$$

□

G Countries with Distinct Decision-Making Weights

Consider an arbitrary active coalition $S \subseteq N$. We introduce country-specific weights, w_i, which are normalized to one in a considered coalition S: $\sum_{i\in S} w_i = 1$. As we are interested in decision making within one considered coalition, we can normalize weights to one in any coalition we analyze. Then, cooperative solution is the result of the minimization of weighted sum of damage functions:

$$\min_{G_S} \sum_{i \in S} w_i D_i$$

The solution reads:

$$k_{wgt} \equiv \sum_{i \in S} w_i k_i$$

where DR^S is the set of drivers: $j \in DR^S$ if $j \in S$ and $k_j > k_{wgt}$, and ND^S is the set of non-drivers: $i \in ND^S$ if $i \in S$ and $k_i \leq k_{wgt}$.

G.1 The Proposition Proof

The Proposition reads:

$$\sum_{j \in DR^S} w_j (D_j(S) - D_j(S \setminus j)) \leq \sum_{i \in ND^S} w_i (D_i(\emptyset) - D_i(S))$$

In the considered specification it takes the following form:

$$\frac{\alpha}{2} \sum_{j \in DR^S} w_j (k_j - k_{wgt})^2 \leq \frac{\alpha}{2} \sum_{i \in ND^S} w_i (k_h - k_{wgt})(k_h + k_{wgt} - 2k_i)$$

Proof Since k_h denotes the largest preferred SG level, the following inequality holds:

$$\sum_{i \in ND^S} w_i k_i \leq \sum_{i \in ND^S} w_i k_h$$

Then we add the term $\sum_{i \in S} w_i k_i$ to both RHS and LHS, arriving at:

$$\sum_{i \in S} w_i k_i + \sum_{i \in ND^S} w_i k_i \leq \sum_{i \in ND^S} w_i k_h + \sum_{i \in S} w_i k_i$$

We then rewrite the LHS and substitute $k_{wgt} \equiv \sum_{i \in S} w_i k_i$ in the RHS:

$$\sum_{j \in DR^S} w_j k_j + 2 \sum_{i \in ND^S} w_i k_i \leq \sum_{i \in ND^S} w_i k_h + k_{wgt}$$

As country-specific weights in a coalition are normalized to one, we may use the following equality $\sum_{i \in DR^S} w_i + \sum_{i \in ND^S} w_i = 1$:

$$\sum_{j \in DR^S} w_j k_j - \sum_{j \in DR^S} w_j k_{wgt} \leq \sum_{i \in ND^S} w_i k_h + \sum_{i \in ND^S} w_i k_{wgt} - 2 \sum_{i \in ND^S} w_i k_i$$

Now we take the sum operator out of brackets:

$$\sum_{j \in DR^S} w_j (k_j - k_{wgt}) \leq \sum_{i \in ND^S} w_i (k_h + k_{wgt} - 2k_i)$$

We then multiply both RHS and LHS by $(k_h - k_{wgt}) > 0$:

$$(k_h - k_{wgt}) \sum_{j \in DR^S} w_j (k_j - k_{wgt}) \leq (k_h - k_{wgt}) \sum_{i \in ND^S} w_i (k_h + k_{wgt} - 2k_i)$$

As $k_h \geq k_j \ \ \forall j \in N$, we have $(k_h - k_{wgt}) \geq (k_j - k_{wgt})$, and therefore:

$$(k_j - k_{wgt}) \sum_{j \in DR^S} w_j (k_j - k_{wgt}) \leq (k_h - k_{wgt}) \sum_{j \in DR^S} w_j (k_j - k_{wgt}) \leq (k_h - k_{wgt}) \sum_{j \in ND^S} w_j (k_h + k_{wgt} - 2k_i)$$

Or just:

$$\sum_{j \in DR^S} w_j (k_j - k_{wgt})^2 \leq (k_h - k_{wgt}) \sum_{j \in ND^S} w_j (k_h + k_{wgt} - 2k_i)$$

Finally, by multiplying both sides by 0.5α we arrive at the expression, which was to be demonstrated:

$$\frac{\alpha}{2} \sum_{j \in DR^S} w_j (k_j - k_{wgt})^2 \leq \frac{\alpha}{2} \sum_{i \in ND^S} w_i (k_h - k_{wgt})(k_h + k_{wgt} - 2k_i)$$

□

H Empirical Calibration

To identify individual preferences of countries, we take the following steps:

1. From the paper by Rickels et. al (2020), we adopt the following estimates:
   $V_i(0)$ - absolute impact on country GVA in the absence of SG deployment
   $V_i(G^*)$ - absolute impact on country GVA in the optimal level of SG deployment
   We then use these data to find the normalized difference:

$$\frac{V_i(G^*) - V_i(0)}{V_i(0)} = V_i$$

2. We use the normalized difference in countries' GVA to estimate the normalized difference in countries' damages when $SG = 0$ and $SG = G^*$ as formulated in our model:

$$2/\alpha \cdot (D_i(0) - D_i(G^*)) = 2G^* k_i - G^{*2} = V_i$$

   . Following the approach in the paper by Rickels et al.,(2020), we define the optimal level of SG (G^*) as GDP-weighted average of countries' preferences $k_i, i \in N$:

$$G^* = \sum_{i \in N} \frac{GDP_i}{\sum_{j \in N} GDP_j} k_i$$

3. We solve the system of equations for each country $i \in N$:

$$2G^* k_i - G^{*2} = V_i$$

4. We put a lower bound of 0 on k_i.
5. Normalize preferred SG levels for all countries to meet the global average of 4 W/m^2.

Acknowledgements The authors are deeply indebted to Robert Stavins for constructive feedback and inspiring conversations. The paper greatly benefitted from valuable comments by Scott Barrett, Johan Eyckmans, Michael Finus, Alton Frye and Dale S. Rothman, and helpful guidance at early stages of this work from Edward Parson. Also, thanks are due to Daniel Heyen and Christian Traeger for their help in improving the clarity of the model presentation. We are grateful to two anonymous referees for their constructive comments, which helped to improve the paper. The authors are responsible for any and all remaining errors.

Funding Author Irina Bakalova was supported by Harvard's Solar Geoengineering Research Program and the Basic Research Program at the HSE University.

Code Availability The code used to obtain the results presented in this paper is available upon the request.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

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(10):1835–1850. https://doi.org/10.1016/S0014-2921(01)00082-4

Belaia M, Moreno-Cruz JB, Keith DW (2021) Optimal climate policy in 3d: mitigation, carbon removal, and solar geoengineering. Clim Change Econ. https://doi.org/10.1142/S2010007821500081

Bunn M (2019) Governance of solar geoengineering: Learning from nuclear regimes. Harvard Project on Climate Change Agreement, pp 51–54

Caparrós A, Finus M (2020) Public good agreements under the weakest-link technology. J Public Econ Theory 22(3):555–582. https://doi.org/10.1111/jpet.12426

Carraro C, Eyckmans J, Finus M (2006) Optimal transfers and participation decisions in international environmental agreements. Rev Int Organ 1(4):379–396. https://doi.org/10.1007/s11558-006-0162-5

Carraro C, Siniscalco D (1993) Strategies for the international protection of the environment. J Public Econ 52(3):309–328. https://doi.org/10.1016/0047-2727(93)90037-T

Chander P, Tulkens H (1992) Theoretical foundations of negotiations and cost sharing in transfrontier pollution problems. Eur Econ Rev 36(2–3):388–399

Cherry TS, Kroll S, McEvoy D, Campoverde D, Moreno-Cruz J (2022) Climate cooperation in the shadow of solar geoengineering: an experimental investigation of the moral hazard conjecture. Environ Polit 1:1–9

Crutzen PJ (2006) Albedo enhancement by stratospheric sulfur injections: A contribution to resolve a policy dilemma? Clim Change 77(3–4):211. https://doi.org/10.1007/s10584-006-9101-y

d'Aspremont C, Jacquemin A, Gabszewicz JJ, Weymark JA (1983) On the stability of collusive price leadership. Can J Econ 16(1):17–25

Diamantoudi E, Sartzetakis ES (2006) Stable international environmental agreements: An analytical approach. J Public Econ Theory 17:1

Eigruber M, Wirl F (2018) Climate engineering in an interconnected world: the role of tariffs. Dyn Games Appl 8(3):573–587

Emmerling J, Tavoni M (2018) Exploration of the interactions between mitigation and solar radiation management in cooperative and non-cooperative international governance settings. Glob Environ Chang 53:244–251

Finus, M., & Furini, F. (2023). The Governance Architecture of Climate Agreements in the Light of Risky Geoengineering. preprint SSRN https://ssrn.com/abstract=4082787.

Finus M, Furini F, Rohrer AV (2021) The efficacy of international environmental agreements when adaptation matters: Nash-Cournot vs Stackelberg leadership. J Environ Econ Manag 109:1

Finus M, McGinty M (2019) The anti-paradox of cooperation: Diversity may pay! J Econ Behav Organ 157:541–559

Heyen D (2016) Strategic Conflicts on the Horizon: R&D Incentives for Environmental Technologies. Clim Change Econ 07(04):1650013. https://doi.org/10.1142/S2010007816500135

Heyen D, Horton J, Moreno-Cruz J (2019) Strategic implications of counter-geoengineering: Clash or cooperation? J Environ Econ Manag 95:153–177. https://doi.org/10.1016/j.jeem.2019.03.005

Heyen D, Lehtomaa J (2021) Solar geoengineering governance: A dynamic framework of farsighted coalition formation. Oxford Open Clim Change 1(1):1. https://doi.org/10.1093/oxfclm/kgab010

Horton JB, Keith DW (2019) Multilateral parametric climate risk insurance: A tool to facilitate agreement about deployment of solar geoengineering? Climate Policy 19(7):820–826. https://doi.org/10.1080/14693062.2019.1607716
Keohane RO, Victor DG (2011) The regime complex for climate change. Perspect Polit 9(1):7–23. https://doi.org/10.1017/S1537592710004068
Lessmann K, Kornek U, Bosetti V, Dellink R, Emmerling J, Eyckmans J, Nagashima M, Weikard H-P, Yang Z (2015) The stability and effectiveness of climate coalitions. Environ Resource Econ 62(4):811–836
Lloyd ID, Oppenheimer M (2014) On the design of an international governance framework for geoengineering. Glob Environ Polit 14(2):45–63
Manoussi V, Xepapadeas A (2017) Cooperation and competition in climate change policies: mitigation and climate engineering when countries are asymmetric. Environ Resource Econ 66(4):605–627
Manoussi V, Xepapadeas A, Emmerling J (2018) Climate engineering under deep uncertainty. J Econ Dyn Control 1(94):207–224
McGinty M (2006) International environmental agreements among asymmetric nations. Oxf Econ Pap 59(1):45–62. https://doi.org/10.1093/oep/gpl028
Mendelsohn R, Dinar A, Williams L (2006) The distributional impact of climate change on rich and poor countries. Environ Dev Econ 11(2):159–178
Meya JN, Kornek U, Lessmann K (2018) How empirical uncertainties influence the stability of climate coalitions. Int Environ Agreements: Polit Law Econ 18:175–198
Millard-Ball A (2012) The tuvalu syndrome. Clim Change 110(3):1047–1066
National Academies of Sciences, Engineering, and Medicine (2021) Reflecting sunlight: recommendations for solar geoengineering research and research governance. The National Academies Press. https://doi.org/10.17226/25762
Nye JS (2019) Notes on Insights from Other Regimes: Cyber. Harvard Project on Climate Change Agreement, pp. 55–59
Parker A, Horton JB, Keith DW (2018) Stopping solar geoengineering through technical means: a preliminary assessment of counter-geoengineering. Earth's Future 6(8):1058–1065. https://doi.org/10.1029/2018EF000864
Parson EA (2014) Climate engineering in global climate governance: implications for participation and linkage. Transnatl Environ Law 3(1):89–110. https://doi.org/10.1017/S2047102513000496
Ricke KL, Moreno-Cruz JB, Caldeira K (2013) Strategic incentives for climate geoengineering coalitions to exclude broad participation. Environ Res Lett 8(1):014021
Rickels W, Quaas MF, Ricke K, Quaas J, Moreno-Cruz J, Smulders S (2020) Who turns the global thermostat and by how much? Energy Econ 91:104852. https://doi.org/10.1016/j.eneco.2020.104852
Schelling TC (1990) The strategy of conflict, 2nd edn. Harvard University Press, Harvard
Smith W (2020) The cost of stratospheric aerosol injection through 2100. Environ Res Lett 15(11):1104. https://doi.org/10.1088/1748-9326/aba7e7
Tol RSJ, Downing TE, Kuik OJ, Smith JB (2004) Distributional aspects of climate change impacts. Glob Environ Chang 14(3):259–272
Urpelainen J (2012) Geoengineering and global warming: a strategic perspective. Int Environ Agreements: Polit Law Econ 12(4):375–389
United Nations Environment Programme (2022). What you need to know about the COP27 Loss and Damage Fund, 29 November. https://www.unep.org/news-and-stories/story/what-you-need-know-about-cop27-loss-and-damage-fund
Weitzman ML (2015) A voting architecture for the governance of free-driver externalities, with application to geoengineering. Scand J Econ 117(4):1049–1068. https://doi.org/10.1111/sjoe.12120
Wilson B (2021) Keynes Goes Nuclear: Thomas Schelling and the Macroeconomic Origins of Strategic Stability. Mod Intellect Hist 18, 171–201.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.