https://economix.fr/

# Mitigation strategies under the threat of solar radiation management

Document de Travail
Working Paper
2019-3

Fabien Prieur
Ingmar Schumacher
Martin Quaas

UMR 7235

EconomiX - UMR7235
Université Paris Nanterre
Bâtiment G - Maurice Allais, 200, Avenue de la République
92001 Nanterre cedex

Email : secretariat@economix.fr

# Mitigation strategies under the threat of solar radiation management*

Fabien Prieur,† Martin Quaas,‡ Ingmar Schumacher§

**Abstract**

The option to tackle climate change by means of Solar Radiation Management (SRM) is mostly thought to reduce efforts of mitigating greenhouse gas emissions. Here we hypothesize that (i) a unilateral threat to employ SRM can induce players to commit to strategies with increased mitigation effort compared to what would be observed at the Nash equilibrium in emission strategies only and (ii) there exists a way to share the burden imposed by commitment to avoid SRM that Pareto dominates an alternative that would involve too high current emission levels then followed by future SRM deployment. To study these hypotheses we develop a two-region, two-stage, two-period game where regions choose mitigation and SRM. While SRM targets regional climate preferences, in line with current scientific evidence its deployment leads to uncertain damages on the other region. We first develop the general theory and then study a more specific linear-quadratic application. Finally we calibrate the model to real-world data and find that hypothesis (ii) holds for plausible values.

**Keywords:** climate change, solar radiation management, heterogeneous damages, strategic interaction, commitment.

**JEL classification:** C72, Q54.

*We thank participants in conferences, seminars and workshops in Aix-en-Provence, Bielefeld, Ecole Polytechnique, IDEFI (Paris), Lisbon, Marseille, Nice, Paris School of Economics and Rennes. This research has benefited from the support of the French National Research Agency program ANR GREEN-econ (ANR-16-CE03-0005).

†EconomiX, Université Paris Nanterre, 200 avenue de la République, 92001 Nanterre Cedex, France. E-mail: fabien.prieur@parisnanterre.fr.

‡Department of Economics, Leipzig University and iDiv, Deutscher Platz 5e, 04103 Leipzig, Germany. E-mail: martin.quaas@idiv.de.

§IPAG Business School, 184 boulevard Saint-Germain, 75006 Paris, France. Tel.: +352 621264575. E-mail: ingmar.schumacher@ipag.fr.

# 1 Motivation

Global carbon emissions have been rising over the last decades and still continue to increase (Le Quéré et al., 2018). The Paris Agreement has, by introducing the Nationally Determined Contributions, changed the previously cooperative climate negotiations to a non-cooperative Nash game. This indicates that coordinated climate negotiations have failed, and the lack of stringent coordinated climate policy is expected to lead to low mitigation efforts and thus substantial warming. Against this background, some in the scientific community argue that solar geoengineering may eventually become an option of last resort (Latham, 1990; Schelling, 1996; Crutzen, 2006; Weitzman, 2009, 2011). Solar geoengineering or solar radiation management (SRM) are technical measures that directly change the earth's radiation balance. One such measure is spraying sulfate aerosol into the stratosphere (Crutzen, 2006).[1] The direct costs of SRM are small, and substantial global cooling could be achieved at low operational costs (Crutzen, 2006; Klepper and Rickels, 2014) such that unilateral deployment of SRM ('free driving') will be possible (Barrett, 2008; Weitzman, 2015). Yet, SRM does not provide a perfect cure for the problem of climate change, as it tackles the symptom and not its cause, which is the continuous increase in greenhouse gas concentration. The cooling effect of SRM is regionally heterogeneous (Ricke et al., 2010; Moreno-Cruz et al., 2012; Quaas et al., 2016), and thus regions have diverging preferences with respect to the amount of SRM deployment (Robock, 2008; Heyen et al., 2015). SRM also affects precipitation patterns in various and often unpredictable ways, with potentially damaging effects in some regions of the world (Allen and Ingram, 2002; Robock, 2008; Ricke et al., 2012; Ferraro et al., 2014; Aswathy et al., 2015). Thus, while it is possible to control one's own climate through SRM, there can be potentially substantial negative climate impacts on other regions of the world.

[1] This is inspired by the eruption of Mount Pinatubo in 1991 in the Philippines. This eruption was accompanied with the release of 20 million metric tons of sulfur dioxide which then reached the upper atmosphere and reacted with water vapor to form a global haze. As a result, scientists observed a decrease in direct solar radiation of about 30% whereas the average temperature on earth dropped by 0.5°C for more than a year.

Due to the lack of experience, the actual risks of the novel technology of SRM are unknown. The literature in climate ethics puts forward reasons against starting SRM in the first place (Gardiner et al., 2010). This includes the argument of a 'slippery slope', according to which starting SRM, even at an experimental scale, would create a dynamic that would lead to large-scale deployment, irrespective of potential consequences (Jamieson, 1996). Also the general public mostly rejects the idea to use SRM for tackling climate change (Merk et al., 2015). One reason why people object against starting to undertake SRM is the widespread opinion that humans should not manipulate nature in the way injecting sulfate would (Merk et al., 2015).

The previous economic literature studying SRM has mostly relied on integrated assessment frameworks (Heutel et al., 2016; Moreno-Cruz and Smulders, 2017; Emmerling and Tavoni, 2018) to analyze how the potential deployment of SRM interferes with efforts of mitigating greenhouse gas emissions under uncertainty; and whether or not to undertake SRM research (Quaas et al., 2017). In this literature a typical finding is that the SRM option reduces the mitigation efforts, as SRM may serve as a substitute for mitigation to tackle catastrophic climate change (referred to as 'moral hazard' in this literature; Keith 2000). As SRM also raises new strategic options, the economic literature has also studied the incentives to unilaterally undertake SRM in a strategic setting (Schelling, 1996; Barrett, 2008; Weitzman, 2015).

In this paper, we are interested in a new strategic interaction that may come out of the availability and affordability of SRM measures. The specific hypothesis we test in our theory is that regions may strategically choose a stronger mitigation effort than in the non-cooperative setting without an SRM option. The idea is that countries may want to remain in the 'safe' domain without geoengineering and thus mitigate strongly enough such that no region will unilaterally deploy SRM. We derive conditions on the regions' preferences, damage functions, abatement technology, and international transfer agreements, under which this hypothesis either holds or does not hold. Additionally, we study if the threat of unilateral SRM deployment may act as a coordination device. We find that regions' interaction may lead to two different equilibria. In the first equilibrium,

regions commit to emission levels such that it is not worthwhile to undertake SRM. In the second one, they continue to release a large amount of emissions knowing that it will trigger SRM deployment in the future. We want to determine under which conditions the former outcome Pareto dominates the latter.

Our contribution is related to Moreno-Cruz (2015), who develops a two-period game where regions non-cooperatively choose their mitigation effort in the first period and then may undertake SRM during the second one. Within this framework, he looks at the impact of unilateral SRM on mitigation efforts at the subgame perfect Nash equilibrium by comparing these efforts with those that would arise at the first best or cooperative solution. He concludes that unilateral SRM deployment (in the future) may induce inefficiently high levels of mitigation by the threatened region when asymmetries in terms of climate change and SRM damages are sufficiently large. He observes that regions "would be better off negotiating a treaty where both countries jointly implement mitigation." However, Moreno-Cruz does not explain how this coordination might take place. This is the contribution of the current paper: our aim is to properly deal with the alternative scenario to SRM deployment. We believe that the analysis should still be conducted in a non cooperative setting and instead involves a commitment problem. Indeed, the main question we want to ask is under which conditions  it is optimal for one or both regions to refrain from emitting too much to avoid the threat of SRM.

To address this issue, we develop a two-region, two-stage, two-period game similar to Moreno-Cruz (2015), where regions face heterogeneous climate damages and have heterogeneous preferences on SRM. In line with the scientific evidence on SRM, we assume the region who employs SRM can control the local climate perfectly. The other region's climate may be disrupted by this use of SRM, and, furthermore, the other region's citizens may have severe concerns about using SRM in the first place. The simplest way to capture this idea is to assume that the use of SRM induces a shift from a certain damage to a (larger) expected damage, especially for the region that is not using SRM. The analysis of the second period problem reveals the existence a threshold level for aggregate emissions that triggers (unilateral) SRM deployment by the region who is the most vulnerable to

climate change. This implies that, as seen from the first period, regions' payoffs are defined piecewise depending on whether first period aggregate emissions exceed or remain below the threshold. Regions – particularly the one who is subject to the threat of SRM – may naturally want to take care of the existence of the threshold because it affects the payoffs. This is where our approach differs from the previous literature, and in particular from Moreno-Cruz (2015). To model this situation, we adopt a commitment perspective. This consists in endowing the regions with another discrete decision. They can choose to commit to constrain aggregate emissions to a level below the threshold beyond which some region would unilaterally deploy SRM. As a special case of this commitment scenario, one region alone may cut emissions to the extent that aggregate emissions remain below the threshold. To form the decision on emissions, the regions have to anticipate the possible outcomes of the interaction in the two next stages (periods).

Our first step is to characterize these equilibria. We find that there are two possible subgame perfect equilibria. Either the regions do not care about the constraint on emissions and choose their emission level freely and non-cooperatively given the future SRM reactions of the regions . This we refer to as the equilibrium without commitment. Or, they may want to take into account the constraint in their first period decision on emissions and then jointly choose emissions such as to comply with the constraint that emissions should be low enough that there will be no SRM deployment in the subgame perfect equilibrium. This characterizes what we call the equilibrium with commitment.

Our second step then is to study the conditions under which one of these equilibria is more likely to emerge as the subgame perfect equilibrium Our results show that even in the worst scenario in which one region alone bears the responsibility of meeting (or not) the constraint on aggregate emissions, this region may prefer to commit to it. This conclusion holds under some conditions that all point to the same requirement, namely that the equilibrium with commitment should not be too painful to this region (compared to the alternative). These conditions involve the cost and benefit of commitment. On the cost side, given that commitment requires to reduce emissions, the threshold should be high enough so that there is an incentive to make the effort. Moreover, the expected damage

under SRM should be high enough compared to the certain damage with commitment. This ensures that the benefit from commitment, in terms of avoided damage, is sizable. Under these conditions, the threat of SRM – by acting as a commitment device – may promote more mitigation than otherwise.

Our last contribution in this article is to empirically investigate if, under empirically reasonable parameter configurations and within this specific setup, we actually observe that the equilibrium with commitment Pareto dominates the equilibrium with unconditional strategies. To study this we calibrate the model to real world data. In addition, we slightly extend the model by allowing for discounting, by introducing region-specific production functions, and by having both regions exposed to potential SRM damages. We find that, even if the region that feels threatened by the potential SRM deployment of the other region is the one that bears the full cost of reducing its emissions, under general parameter configurations the equilibrium with commitment would still be the Pareto dominating equilibrium. We interpret this as indication that the coordination strategy may be a superior strategy (compared to the non-cooperative one) after all.

The paper is organized as follows. Section 2 develops the general game-theoretic set up and characterizes candidates for subgame perfect Nash equilibria in SRM and emissions. In Section 3, we develop a simple linear quadratic game which allows us to address the main question raised by the present analysis. In Section 3.2 we calibrate the linear quadratic model to real world data and show that indeed the commitment outcome may be preferred for plausible parameter values. Section 4 concludes.

# 2 General game-theoretic set up

We split the world into two regions,[2] which are asymmetrically impacted by climate change. We furthermore restrict our focus to two periods, which is sufficient to capture the essential intertemporal trade-offs that arise in this model. We assume that, during

[2]This allows us to keep a more compact notation but it introduces some requirement for coordination that we do not model in this paper. We will discuss the implication later.

the first period, our two regions, indexed by $j = i, -i$, release emissions, $e_j$, for a private benefit, $F(e_j)$, with $F'(e_j) > 0$, $F''(e_j) < 0$. The resulting aggregate emissions $e = \sum_j e_j$, which form the stock of greenhouse gases in period 2, affect temperature, which in turn determines the extent of the climate damage that will be felt in the second period. As the relationship between cumulative emissions and the global mean surface temperature is linear (IPCC, 2013), we can choose units of measurement such that temperature equals the stock of greenhouse gases (Moreno-Cruz and Smulders, 2017). In the second period, regions decide over SRM measures, of an extent $g_j \geq 0$, which come at a monetary cost $C(g_j)$, with $C(0) = 0$, $C(g_j)' > 0$, $C''(g_j) > 0$. The potential corner solution $g_j = 0$ is of particular interest. Whereas we assume that the two regions are symmetric with respect to private benefits of emissions and costs of SRM – we take this assumption simply for mathematical convenience and relax it later – they are heterogeneous with respect to the climate damages which materialize in period 2.

We consider SRM as a last resort option to be used once severe climate damages are expected to occur. If no region chooses to deploy SRM, then both regions suffer from a *simple*, asymmetric climate damage, $D_j^c(e)$, with $D_j^{c\prime}(e) > 0$, $D_j^{c\prime}(e) > 0$. In the absence of SRM, region $-i$ is more vulnerable to climate change than region $i$, i.e. marginal damages are higher, $D_i^{c\prime}(e) < D_{-i}^{c\prime}(e)$, $\forall e$.

Our first departure from standard modeling is that we assume that as soon as a region undertakes SRM, then the system shifts to a different regime. This different regime is characterized by a new source of uncertainty that is due to the implementation of SRM. We suppose that once a region undertakes SRM, then the uncertainty surrounding the impact of SRM on the climate system and the beliefs held by citizens about the acceptability of SRM makes damage essentially uncertain.

We model the system's shift to a new risky world by a change in the damage function from $D_j^c(e)$ to $D_j^s(\tilde{e}, \omega)$, with $\tilde{e} = e - g$, $g = \sum_j g_j$, and $\omega$ a discrete random variable taking values in the set $\Omega = \{\underline{\omega}, \overline{\omega}\}$, with well-defined probability distribution. Thus, the immediate benefit of SRM is that the temperature is reduced to the 'engineered' temperature $\tilde{e}$, while the indirect cost (in addition to the direct cost of implementing

SRM) is the uncertainty it introduces over the climate system. Let us denote $\mathbb{E}[D_j(\tilde{e})]$ as the expected damage. If engineered temperature $\tilde{e}$ was a the same level as $e$, $\tilde{e} = e = x$, both absolute and marginal damages would be higher, i.e.

$$D_i^c(x) \leq \mathbb{E}[D_i(x)] < \mathbb{E}[D_{-i}(x)], \quad \forall x. \tag{1a}$$

$$D_i^{c\prime}(x) \leq \mathbb{E}[D_i'(x)] < \mathbb{E}[D_{-i}'(x)], \quad \forall x. \tag{1b}$$

This means that also with SRM, region $-i$ remains more vulnerable to climate change than region $i$. The direct implication of these assumptions is that region $-i$ is the one with the highest incentive to deploy SRM. Thus, if there is one region that may want to undertake SRM, then it will be region $-i$. Given that emissions in period 1 affect temperature in period 2, and thereby the incentives to undertake SRM, then it is clear that the decision to undertake SRM will be determined by the stock of greenhouse gases inherited from the first period. Thus, we may expect that if cumulative emissions remain very low, then there will be no benefit from deploying SRM. Furthermore, since SRM involves the shift to (uncertain) climate damages, this information should matter to the regions when they choose their emission levels. Hence the timing of the game is crucial.

## 2.1 Timing and strategies of the game

Our problem can suitably be designed as a two-stage, two-period game by the two regions:

1. First stage (commitment problem): The regions choose whether or not commit to reducing emissions to jointly meet the threshold that there will be no SRM in the future.

2. Second stage (two-period game):

   2.1 Choice of the emission levels,

   2.2 Choice of the (non-negative) amount of SRM.

The second stage comprises two periods that form the climate change game in which continuous decisions, regarding emissions and SRM deployment, have to be taken. We

begin with a presentation of the timing and strategies in this second stage, before moving back to the first stage where only one discrete choice remains.

In the first period, regions non-cooperatively choose their emission levels, whereas in the second period, they will have to decide on the levels of SRM deployment. Both decisions are interwoven. Depending on the level of aggregate emissions in period 1, the Nash equilibrium in period 2 may or may not include positive levels of SRM. Clearly, there will be some critical thresholds in terms of aggregate emissions such that none, only one, or both regions will find it optimal to deploy SRM. This also implies that in some cases, these thresholds may enter each region's decision problem as an additional constraint in the first period.

Regions should take into account this information when choosing their emissions levels in period 1. As to the strategic choice of emissions, we then expect that two types of interaction may occur. Either the regions get rid of the (potential) constraint on aggregate emissions and freely choose their emission levels in the first period, being aware that this will trigger SRM deployment in period 2. This is the first candidate to the subgame perfect equilibrium (SPE) that we will refer to as the equilibrium in *unconditional strategies* (US). Or, they decide to restrict aggregate emissions to a level such that SRM deployment will not occur in period 2. This provides us with a second candidate to the SPE that we label the equilibrium in *conditional strategies* (CS).

From the analysis of the second stage, there possibly exist two SPE, the CS and the US. This means that besides their continuous decisions, players have a discrete choice to make. This brings us to the first stage of the game in which regions have to decide which equilibrium to play. Here it is worth emphasizing that since the CS equilibrium naturally comes with a reduction in aggregate emissions (compared to the US), it raises the question of how the resulting burden is split among the two regions and involves some sort of *coordination* and *commitment*. We do not model the negotiation process that would lead to a particular sharing rule as our aim is to focus on the commitment problem. We consider that regions have to commit to a costly reduction in emissions based on a *given sharing rule*, and rely on an individual rationality condition (participation

constraint). This means that commitment and the CS equilibrium will arise if and only if no region obtains a reduction in its welfare compared to what it would get by playing the US equilibrium.

In the end, the question we seek to address is the following: Is there a way to share the burden imposed by the reduction of emissions that comes with commitment compatible with both regions not being worse off in the conditional scenario (CS) than in the unconditional scenario (US)? To answer this question, the problem is solved backwards, starting with the second period (of the second stage) problem.

Finally note that we can define a reference scenario as the one where regions non-cooperatively choose emissions in the first period, but cannot invest in SRM in the second period. This is a simultaneous game in emissions where emissions are unconstrained by the potential investment in SRM in the second period. Hence we call the equilibrium that results from this reference scenario the *unconstrained Nash* equilibrium (UN). This will be helpful to compare to the two equilibria discussed above.[3]

## 2.2 Second period: SRM decisions

We proceed backwards for the resolution starting with the SRM game of the second period where the stock of greenhouse gases – i.e, first period's emissions – is taken as given. Given the stock of greenhouse gases, $e$, and the other region's strategy, region $j$ solves:

$$\min_{g_j \geq 0} \left\{ \mathbb{E}[D_j(e-g)] + C(g_j) \right\}.$$

The first order condition (FOC) is given by:

$$\begin{cases} \mathbb{E}[D_j'(e-g)] - C'(g_j) \leq 0, \\ \left(\mathbb{E}[D_j'(e-g)] - C'(g_j)\right) g_j = 0, \quad g_j \geq 0 \text{ for } j \in \{i, -i\}. \end{cases}$$

[3]The UN can actually be a SPE of the two-period game. However, this occurs only when resulting aggregate emissions are too low to induce future SRM deployment and/or SRM deployment is too costly for the US equilibrium to exist. In the coming analysis, we discard this case because it is not the interesting one.

These first order conditions allow us to identify two thresholds for the stock of greenhouse gases, $0 < \bar{e}_{-i} < \bar{e}_i$, that trigger SRM deployment by respectively player $-i$ and $i$. Depending on the level of cumulative emissions in period 1 there may exist three (mutually exclusive) Nash equilibria in period 2:

1. a corner equilibrium without SRM, $g = 0$, iff $e \in [0, \bar{e}_{-i}]$,
2. an equilibrium with unilateral SRM by region $-i$, $g_{-i} > 0$ and $g_i = 0$, iff $e \in (\bar{e}_{-i}, \bar{e}_i]$,
3. an interior equilibrium with positive SRM by both regions, $g_{-i} > g_i > 0$, iff $e > \bar{e}_i$.

Not surprisingly, we find a corner solution if the stock of greenhouse gases is very low, an interior solution if it is very high, and there is an intermediate situation in which region $-i$ is the only one to undertake SRM. We are particularly interested in the threshold $\bar{e}_{-i}$ between no SRM and unilateral SRM deployment by region $-i$. With a slight abuse of notation, and omitting for now the index, this threshold $\bar{e}$ is implicitly defined by $C'(0) = \mathbb{E}[D'_{-i}(\bar{e})]$ and thus explicitly given by $\bar{e} = (\mathbb{E}[D_{-i}]')^{-1}(C'(0))$. For all $e > \bar{e}$, it is then possible to characterize a reaction function:

$$g_{-i} = g_{-i}(e) \text{ with } g'_{-i} = \frac{\mathbb{E}[D''_{-i}]}{C'' + \mathbb{E}[D''_{-i}]} \in (0, 1). \tag{2}$$

The level of the threshold $\bar{e}$ will be crucial in the analysis of the first period's outcome.

## 2.3 First period: emission decisions

We now look at the emission decisions in the first period. First we introduce the UN equilibrium, which is our reference game where regions are not allowed to invest in SRM. Thus, regions do not need to take into account how their emissions may influence the second period's strategies. We then study how emissions in unconditional strategies are influenced by their impact on the potential SRM in the second period. This yields an equilibrium that will be one candidate for the SPE. Then we look at the game that would take place under commitment where emissions are chosen with strategies conditional on

each other and the second period's SRM choices. This will yield our second equilibrium that will be another candidate for the SPE.

### 2.3.1 Unconstrained Nash equilibrium

We first study the reference game where both regions do not have the option to undertake SRM and simultaneously choose emissions. We refer to the outcome as the UN equilibrium. Aggregate emissions are denoted by $e^u = e^u_{-i} + e^u_i$, where the two regions choose their individual emission levels such as to maximize their net payoffs, taking the other region's emissions as given. The individual emission levels are implicitly given by

$$F'(e^u_j) = D^{c\prime}_j(e^u) \text{ for } j = i, -i. \tag{3}$$

Thus, aggregate emissions, $e^u$, are given by

$$e^u = M(e^u) \text{ with } M(e^u) = \sum_j (F')^{-1}(D^{c\prime}_j(e^u)) \text{ and } M'(e^u) = \frac{\sum_j D^{c\prime\prime}_j}{F''} < 0.$$

If aggregate emissions were less than the threshold $\bar{e}$ identified above, then no region would want to undertake SRM in the second period. As this is a somewhat trivial solution we now impose

$$\bar{e} < e^u. \tag{Condition 1}$$

Given Condition 1, (at least) region $-i$ would have an incentive to undertake SRM in the second period if both regions were to emit as if the SRM option was not available.

Note, however, that the UN remains a candidate for the SPE when the SRM option is available and Condition 1 holds. Indeed, in this context, region $-i$ will undertake SRM if and only if it expects a higher payoff with SRM than without (that is at the UN). In the coming analysis, we will focus on the situation in which SRM by region $-i$ represents a credible threat, meaning that SRM deployment is a dominant strategy and the UN is no longer a possible SPE.

### 2.3.2 Emissions in unconditional strategies

Suppose that regions choose their emissions without staying below the threshold that will trigger SRM deployment by region $-i$.[4] Both regions anticipate the SRM deployment by region $-i$ and the effect on expected climate damages. The two regions choose emission levels $e_i$ and $e_{-i}$ according to

$$\begin{aligned} e_i^n &= \arg\max_{e_i} \left\{ F(e_i) - \mathbb{E}[D_i(e_i + e_{-i} - g_{-i}(e_i + e_{-i}))] \right\}, \\ e_{-i}^n &= \arg\max_{e_{-i}} \left\{ F(e_{-i}) - \mathbb{E}[D_{-i}(e_i + e_{-i} - g_{-i}(e_i + e_{-i}))] - C(g_{-i}(e_i + e_{-i})) \right\}, \end{aligned} \tag{4}$$

where both players also anticipate that region $-i$ will choose the amount of SRM in period 2 such as to maximize payoffs for the given stock of greenhouse gases. The two first-order conditions that characterize the resulting Nash equilibrium thus become, denoting $e^n = e_i^n + e_{-i}^n$,

$$\begin{aligned} F'(e_i^n) &= (1 - g'_{-i}(e^n))\mathbb{E}[D'_i(e^n - g_{-i}(e^n))] \\ F'(e_{-i}^n) &= \mathbb{E}[D'_{-i}(e^n - g_{-i}(e^n))]. \end{aligned} \tag{5}$$

These conditions can be compared to the ones characterizing the UN, given in equations (3). For region $-i$, the marginal cost of emissions (at the right-hand side) is higher under SRM because of the switch from a certain to an expected damage. This is the *hazard effect* of SRM deployment on the emission trade-off. For region $i$, a similar effect is present, but in addition, the condition includes an additional term $(1 - g'_{-i}(.))$ that is multiplied with expected marginal climate damage. This is because region $i$ takes into account region $-i$'s reaction to an increase in its emissions. It knows that part of its emissions will be offset by SRM, which tends to lower the marginal cost of emissions. We dub this the *cooling effect*.

Both effects work in opposite directions on the emissions of player $i$. If the cooling effect is very strong, then the additional uncertainty arising from SRM deployment is less

[4] We could easily extend the subsequent analysis to one with many regions. However, to avoid mathematical clutter we decided to opt for two regions alone, as this is sufficient to provide the general idea of the theory that we develop here.

important as region $i$ anticipates that region $-i$ will tackle it through SRM. However, if the hazard effect sufficiently outweighs the cooling effect, then SRM works as a threat to region $i$ as it then mainly increases the damage costs for region $i$. Overall, the hazard effect dominates over the cooling effect if the following condition holds:

$$(1 - g'_{-i}(y))\mathbb{E}[D'_i(x)] \geq D_i^{c\prime}(x) \text{ for all } x, \qquad \text{(Condition 2)}$$

with $x = \tilde{e} = e$ and $y$ implicitly defined by $y = x + g(y)$. The LHS of the inequality represents the expected marginal climate damage under SRM. Note that this is a stronger condition than the ranking imposed in Condition 1. However, only if this condition is satisfied would region $i$ want to act qualitatively different when faced with the threat of SRM deployment.

Overall, solving for this problem, we get a first candidate for the subgame perfect equilibrium defined by the pair of strategy profiles $(\{e_i^n, 0\}, \{e_{-i}^n, g_{-i}(e^n)\})$.

### 2.3.3 Emissions in conditional strategies

Suppose that both regions choose their emissions taking the other region's emissions as given, but conditional on the constraint that aggregate emissions do not exceed the threshold level $\bar{e}$. Then, they face a common constraint, $e \leq \bar{e}$, and solve:

$$\max_{e_j} F(e_j) - D_j^c(e) \text{ subject to } e \leq \bar{e} \text{ for } j \in \{i, -i\}.$$

Under Condition 1, the constraint is binding, i.e. $e = \bar{e}$, and the FOCs are given by:

$$F'(e_j) = D_j^{c\prime}(e) + \lambda_j, \tag{6}$$

with $\lambda_j \geq 0$ the (marginal) cost, for player $j$, of meeting the constraint.

Because of the common constraint that couples regions' strategy spaces to each other, we use a specific equilibrium concept, the *coupled constraint Nash equilibrium*, as introduced by Rosen (1965). Following Krawczyk (2005), we can introduce a pair of weights $(\varepsilon_i, \varepsilon_{-i}) \in [0,1] \times [0,1]$ such that $\sum_j \varepsilon_j = 1$ and $\lambda_j = \varepsilon_j \lambda$, with $\lambda > 0$, the social value

of meeting the constraint. Then these weights $\varepsilon_j$ represent region $j$'s 'responsibility' in meeting the constraint. In other words, the weights give the share of the burden imposed by the constraint. Take $\varepsilon = \varepsilon_i$, such that $\varepsilon_{-i} = 1 - \varepsilon$. In the remainder of the analysis we will refer to $\varepsilon \in [0, 1]$, as a 'sharing rule'.

Existence and uniqueness of a coupled constraint equilibrium, for any given sharing rule $\varepsilon$, follows straightforwardly from our assumptions above. In this equilibrium, emission levels will be defined as a function of the sharing rule (thus the superscript $r$), $e_j^r(\varepsilon)$ for $j \in \{i, -i\}$, and aggregate emissions will coincide with the threshold $\bar{e}$. This equilibrium in conditional strategies can then be described by the pair of strategies $\left(\{e_i^r(\varepsilon), 0\}, \left\{e^r_{-i}(\varepsilon)\right), 0\right\}$.

## 2.4 Commitment problem

What comes out of the previous analysis is the existence of two candidates to the subgame perfect Nash equilibrium (SPE). The first possible SPE arises from the non-cooperative strategies of both regions, which may lead to an equilibrium where either no region, one region or both regions invest in SRM. Under Condition 1 we know that at least one player wants to undertake SRM in the second period, while under Condition 2 we know that only region $-i$ has incentives to undertake SRM. The second possible SPE is the one that follows when both regions comply with the threshold $\bar{e}$, but still choose their emissions non-cooperatively, in order to remain in the regime without SRM. The question we now ask is under which conditions does the CS equilibrium arise as the SPE?

We will not model the negotiation process that yields us a specific sharing rule, $\varepsilon$. The literature has identified sufficiently many of these and, while we know that different negotiation processes lead to different sharing rules, it is enough for our purpose to characterize the resulting equilibria for a given sharing rule. We thus study which set(s) of sharing rules gives rise to the different equilibria, rather than asking what comes out of a particular negotiation process. Hence, we build the analysis on an individual rationality condition according to which regions choose to obey the condition of not exceeding the

threshold (and then play CS) if they expect a payoff that is not lower than what they would get under US. The answer to the above question ultimately involves the comparison between payoffs at the SPE candidates. We then define $\Pi_j^n$ and $\Pi_j^r(\varepsilon)$ as the payoffs at the US *vs.* CS, and denote by $\Pi_j^u$ the payoffs at the UNE.

As for CS, we especially look at two sharing rules. First, we deal with the case $\varepsilon = 1$, in which region $i$ bears the whole responsibility of complying with the constraint, and search for the conditions under which $\Pi_j^n \leq \Pi_j^r(1)$ for $j \in \{i, -i\}$. This is clearly the worst case for region $i$. Yet, it is natural to look at this case since region $-i$ is the one that undertakes SRM and given that it this imposes a threat on region $i$, the latter may unilaterally find it optimal to lower its emissions. On the other hand, one may argue that this situation, where one region bears all the burden of meeting the constraint $e \leq \bar{e}$, is not relevant because at the international stage there is often some kind of multilateralism (Horton, 2011). Hence we also study the general case in which regions share the burden, $\varepsilon < 1$, and highlight the conditions for $\Pi_i^n < \Pi_i^r(\varepsilon)$ for $j \in \{i, -i\}$.

As our general model does not allow us to derive specific conditions, in the remainder we consider a linear quadratic application that allows for a full analytical treatment.

# 3 Linear quadratic application

We now turn to a linear-quadratic specification of functional forms, which allows us to characterize and derive closed-form solutions for the two candidates for the subgame perfect equilibrium (SPE) and then identify under which sharing rules either of the two is the actual SPE. All calculations are relegated into the Appendix.

We specify the functional forms as follows

$$F(e_j) = a\, e_j \left( b - \frac{1}{2}\, e_j \right) \tag{7a}$$

$$D_j^c(e) = \frac{d_j}{2} e^2 \tag{7b}$$

$$C(g_j) = c_1\, g_j + \frac{c_2}{2}\, g_j^2 \tag{7c}$$

They satisfy all assumptions made above. Moreover we choose the simplest description of the impacts of SRM. We consider that under SRM the damage function becomes $D_j^s = \omega D_j^c(e)$, with $\omega$ a discrete random variable. For instance, we may assume that there exist two states of the world: $\underline{\omega} = 1$, with probability $p_j$, and $\overline{\omega} > 1$, with probability $1 - p_j$. Denoting $\tilde{d}_j$ as the expected value of the slope of the marginal damage function for region $j$, for simplicity we further impose

$$d_i \leq \tilde{d}_i < d_{-i} = \tilde{d}_{-i}, \tag{8}$$

which boils down to assuming that region $-i$ perfectly controls the impact of SRM on the local climate ($p_{-i} = 1$), but SRM represents a potential threat to region $i$.[5]

In this example, Condition 1 can be written as

$$e^u = \frac{2\,a\,b}{a + d_i + d_{-i}} > \bar{e} = \frac{c_1}{d_{-i}} \tag{9}$$

which yields a necessary and sufficient condition on parameter values for SRM deployment in the second period. Condition 2 can be written as the following condition on parameter values:

$$\tilde{d}_i \geq d_i \left(1 + \frac{d_{-i}}{c_2}\right) \equiv \underline{\Delta}_i \tag{10}$$

This condition states that the marginal damage under SRM needs to be high enough, such that player $i$ would like to prevent SRM from happening. The right-hand-side of Condition (10) defines the threshold value for this.

Next, we study the existence of each possible SPE, which boils down to showing that the different emission subgames in the first period of the second stage have a solution.

## 3.1 Candidates for the subgame perfect equilibrium

The following proposition summarizes the existence conditions for the UN equilibrium and the two different candidates for the SPE we are interested in, the US equilibrium where regions non-cooperatively choose their emission levels in the first period, being aware

[5]Relaxing this assumption would not change our qualitative results.

of the shift into an SRM regime in period 2, and the CS equilibrium where aggregate emissions in the first period are chosen such that there will be no SRM in the second period.

**Proposition 1** *Under conditions (8)—(10), there exists a unique*

*a) UNE featuring* $e_j^u \geq 0$ *for all* $j$ *if and only if*

$$a \geq d_{-i} - d_i. \tag{11}$$

*b) US equilibrium with* $e_j^n \geq 0$ *for all* $j$ *and* $e^n > \bar{e}$ *if and only if*

$$\tilde{d}_i < \frac{c_2 + d_{-i}}{c_1\, c_2}\, (2\, a\, b\, d_{-i} - c_1(a + d_{-i})) \equiv \overline{\Delta}_i. \tag{12}$$

*c) CS equilibrium conditional on* $\varepsilon$ *with* $e_j^r(\varepsilon) \geq 0$ *for all* $j$ *and for all* $\varepsilon$ *if and only if*

$$a\, b\, d_{-i} - c_1\, (a + d_i) \leq 0. \tag{13}$$

**Proof 1** *See the Appendix A.*

Conditions (11) and (13) ensure that individual emissions are non-negative at the UN and the CS. Condition (11) tells us that the heterogeneity in terms of damage must be bounded from above in order for both regions to release a non-negative amount of emissions at the Nash equilibrium. Condition (13) ensures that, whatever the scenario, no region has the capacity to saturate the constraint on aggregate emissions imposed by the threshold with its own emissions. This is somehow very demanding but allows us to avoid unnecessary complications.[6]

According to condition (12), for the aggregate emissions at the UN to effectively lie above the threshold, actual marginal damage should not be too high for this region to ignore the constraint and play the no commitment equilibrium.[7] Overall, the number of

[6]The alternative is to define the interval of variation of $\varepsilon$ compatible with non-negative emissions.

[7]Condition (9) guarantees that the right-hand-side of (12) is positive.

existence conditions and assumptions at this stage is quite limited. Besides the conditions required to meet the non-negativity constraints, the two additional conditions (10) and (12) define an interval $[\underline{\Delta}_i, \overline{\Delta}_i]$ of variation for $\tilde{d}_i$, with boundaries given by the right-hand sides of Conditions (10) and (12). As should already be clear, this parameter $\tilde{d}_i$ will play a crucial role in the coming analysis.

Last, we obtain that aggregate emissions are lower at the CS than at the US (and the UN). But the comparison between individual emissions is not straightforward. At least we can conclude in the limit cases: when all the burden imposed by the constraint is on region $i$'s shoulders, i.e., $\varepsilon = 1$, its emissions will be set below the ones released at the US. So this region will reduce emissions even further compared to the benchmark (UN). This allows region $-i$ to increase its own emissions ($e^r_{-i}(1) > e^n_{-i}$). We get the opposite for $\varepsilon = 0$.

We now search for the conditions under which commitment emerges in the first stage.

### 3.1.1 Comparison between regions' payoffs

First suppose that the burden imposed by the constraint on aggregate emissions lies on region $i$'s shoulders only: $\varepsilon = 1$. As region $-i$ does not incur costs of SRM, it prefers the CS in which region $i$ only takes care of the constraint over US. So we have to compare the payoffs obtained by region $i$ in the US *vs.* CS equilibrium. These payoffs are respectively given by (multiplied by $2a$; see Appendix):

$$\Pi^r_i(1) = \left(-3\,(a\,b)^2 + 4\,a\,b\,(a + d_{-i})\bar{e} - ((a + d_{-i})^2 + a d_i)\bar{e}^2\right), \tag{14}$$

$$\Pi^n_i = (ab)^2 - \tilde{d}_i \left(a + \left(\frac{c_2}{c_2 + d_i}\right)^2 \tilde{d}_i\right) (\tilde{e}^n)^2, \tag{15}$$

where $\tilde{e}^n$ is the 'engineered' temperature in the US equilibrium, which is independent of $d_i$, as region $-i$ chooses the engineered temperature according to its preferences. Also $\bar{e} = c_1/d_{-i}$ does not depend on $\tilde{d}_i$.

We get the following result:

**Proposition 2** *There exists a unique* $\hat{\Delta}_i < \overline{\Delta}_i$ *such that* $\Pi_i^r(1) \geq \Pi_i^n$ *for all* $\tilde{d}_i \geq \hat{\Delta}_i$.

In other words, region $i$ finds it optimal to commit to meeting the threshold in order to avoid SRM deployment if and only if the relative benefit from committing is sizable enough, with takes the form of the avoided extra damage from SRM and is captured by the size of $\tilde{d}_i$.

If the condition $\tilde{d}_i > \hat{\Delta}_i$ is not met in practice, there still may be scope for sharing the burden of meeting the constraint in such a way that both regions are better off than in the US equilibrium. We thus now study whether there are conditions under which $\Pi_j^n < \Pi_j^r(\varepsilon)$ for both regions $j = i, -i$ even if $\tilde{d}_i < \hat{\Delta}_i$. Compared to the literature applying the concept of coupled constraint Nash equilibrium to environmental problems (see among others Tidball and Zaccour 2009; Morgan and Prieur 2013), we do not impose a particular sharing rule, $\varepsilon$. Rather our aim is to study if there exist rules that are compatible with commitment at the equilibrium.

For that purpose, we consider the most interesting situation where region $-i$ would prefer the US to the CS where it bears all the constraint ($\Pi_{-i}^r(0) < \Pi_{-i}^n$). In this case, from the direct comparison of the two regions' payoffs in the US and the CS equilibrium (see Appendices A.3.1 and A.3.2), it is possible to determine a critical sharing rule for each region according to which they prefer the CS equilibrium. Clearly, the larger $\varepsilon$, the better region $-i$, and the worse region $i$. Let us denote these critical sharing rules respectively as $\overline{\varepsilon}_i$ and $\underline{\varepsilon}_{-i}$. A sharing rule where both regions are better off under CS than under US exists if and only if $\underline{\varepsilon}_{-i} \leqq \overline{\varepsilon}_i$. The remaining open question thus is under which conditions this inequality holds true. This leads us to the following Proposition.

**Proposition 3** *If* $\tilde{d}_i$ *is larger than a critical threshold* $\tilde{\Delta}_i < \hat{\Delta}_i$ *and*

$$\frac{\bar{e}}{e^u} > \frac{(a + d_i + d_{-i})^2 - 4d_{-i}(a + d_{-i})}{(a + d_i + d_{-i})^2 - 4d_i\, d_{-i}}, \tag{16}$$

*then there exists a non-empty range of variation of* $\varepsilon$*, that includes the uniform rule* $\varepsilon^u = \frac{1}{2}$*, such that for every sharing rule taken in this interval both regions find it optimal to commit.*

| | | | | |
|---|---|---|---|---|
| region $i$ would benefit from SRM | both US and CS are SPE | some sharing rule $0 < \varepsilon < 1$ exists such that both regions prefer CS over US | region $i$ prefers CS over US even if it has to bear the whole cost, $\varepsilon = 1$ | player $-i$ would not undertake SRM in US |
| cooling effect dominates hazard effect | | $\exists \varepsilon \in [0,1]$ such that $\Pi_i^r(\varepsilon) > \Pi_i^n$ and $\Pi_{-i}^r(\varepsilon) > \Pi_{-i}^n$ | $\Pi_i^r(1) > \Pi_i^n$ | US equilibrium implies $e^n < \overline{e}$ |

$\underline{\Delta}_i$ $\tilde{\Delta}_i$ $\hat{\Delta}_i$ $\overline{\Delta}_i$

increasing marginal damage $\tilde{d}_i$ under SRM in region $i$

Figure 1: Illustration of the different cases with respect to the marginal damage under SRM in region $i$, $\tilde{d}_i$.

In words, the threat of SRM may induce regions to commit to not exceeding a critical emission threshold, thereby increasing global mitigation compared to the equilibrium with unconditional strategies. Of course, commitment is more likely to occur when regions share the burden imposed by the reduction of emissions. The conditions under which this conclusion holds have to do with both the cost and benefit of commitment. The benefit takes the form of an avoided extra damage induced by SRM, and it should be sizable for the region that faces the threat of SRM. In other words, the expected damage under SRM should be high enough. As to the cost, commitment comes with a cost since it forces regions to reduce their emissions compared to the US scenario. Clearly, the larger the threshold, the lower the effort and the higher the incentive to commit. This is the meaning of condition (16).

Overall, we get the opposite conclusion compared to the literature that adopts a centralized perspective and emphasizes that the option to undertake SRM in the future should induce regions to reduce their current mitigation efforts (Jamieson, 1996; Keith, 2000; Quaas et al., 2017). The difference can be explained by the role of strategic interactions and the interpretation of SRM as a commitment device.

## 3.2 Numerical application

In order to quantify the effects and to see whether the equilibrium in conditional strategies may be the one that actually comes out of the interaction we calibrate the linear quadratic model to real world data. For that we have to calibrate the temperature-emission function, the damage function, the cost of SRM and the production function.

### 3.2.1 Calibration

Our calibration of the emissions-carbon-temperature relationship is based on the IPCC A2 scenario, according to which the relationship between emissions and the carbon stock is approximately linear. Similarly, as the AR5 IPCC Synthesis report shows, the relationship between the carbon stock and temperature change is approximately linear, too. We can thus assume that the temperature-emission relationship is given by

$$T_{t+1} = T_t + \gamma(e_{it} + e_{-it}).$$

Denoting the change in temperature by $dT = T_{t+1} - T_t$ we then get

$$dT = \gamma\,(e_i + e_{-i}).$$

Based on the simulations underlying the A2 climate scenario we obtain $dT = 0$ with $(e_{it} + e_{-it}) = 0$ (preindustrial); $dT = 1°\text{C}$ with $(e_{it} + e_{-it}) = 8.6$ (GtC per yr in 2010); $dT = 4.5°\text{C}$ with $(e_{it} + e_{-it}) = 28$ (GtC per yr in 2100 based on RCP 8.5). Assuming our linear relationship between emissions and the change in temperature implies thus $\gamma = .12$, or $\gamma = .16$. We choose an average given by $\gamma = 0.14°\text{C/GtC}$.

We calibrate the damage-temperature relationship based on country-level, climate change impact data from Roson and Sartori (2016). They estimate the country-level damages in percent of GDP from a 3°C warming. As the database does not include all countries, we extrapolate impacts for those missing observations from neighboring countries. The GDP data we take from the SSP1 scenario, provided at the country-level, for the year 2030. Thus the assumption is that the emissions from 2010 induce damages

in 2030, which also gives rise to the recursive feature of the model. We assume this is the appropriate delay in the carbon cycle between emissions and their maximum impact on temperature.

We calibrate damages according to the quadratic form

$$\psi_i \cdot \text{GDP}_i = \frac{d_i}{2} (dT)^2,$$

where $\psi_i$ is the country-specific damage, $\text{GDP}_i$ is GDP of country $i$ in the year 2030, and $d_i$ is the percent GDP damage coefficient. As the $\psi_i$ data is given for a 3°C warming, then for $dT = 3$°C we can solve for country-level GDP impacts, at a generic temperature level, $d_i$. Thus, our damage function is

$$D_i^c(dT) = \frac{d_i}{2} \cdot dT^2.$$

Considering the simplest case, that is a binary random variable with possible realizations $\underline{\omega} = 1$ and $\overline{\omega} > 1$, the expected damage parameter is $(p_i\, d_i + (1 - p_i)\, \bar{d}_i)$ with $\bar{d}_i = \overline{\omega} d_i$. We have no data about the different probabilities or the country-specific expected changes to the temperature-damage relationship that arise from the use of SRM. Given this uncertainty, we undertake sensitivity analysis with respect to $\tilde{d}_i$. For the baseline calibration we choose $p_i = 0.3$, $\bar{d}_i = 1.1 \cdot d_i$, $\bar{d}_{-i} = d_{-i}$.

As we calibrate the damages to 2030, we also generalize our theoretical model by allowing for discounting. We assume an annual discount rate of 5% which we attach to the damages, and denote the corresponding discount factor by $\beta$.

For the costs of applying an amount $g_i$ of SRM, we consider the functional form specified in equation (7c). Moriyama et al. (2017) argue that the marginal cost for SRM based on presently existing technology is somewhere around 90 billion USD/$W/m^2$. In order to use this estimate for $c_1$, we need to transform it into temperature units. We know that $3.7 W/m^2 = 2$°C warming, and thus one degree warming requires $3.7/2 = 1.85$ $\text{W}/m^2$. Thus we obtain $c_1 = 0.09 \cdot 1.85$ trillion USD per degree cooling. For $c_2$ we have no information, so we shall do sensitivity analysis on this. In the baseline calibration we set $c_2 = 0.1$.

Our data consists of emissions and GDP for each country in the world. In order to stick to the model, we aggregate these country-level observations into two regions. We relax the assumption of the analytical model that benefits of emissions are identical across regions, but we maintain the assumption of the linear quadratic specification of functional form (eq. 7a). Having these two regions, we thus need to estimate two parameters per production function. Our approach is to calibrate the model under the assumption that currently the countries behave according to the UN strategies, i.e., without taking the SRM option into account. We then solve for the coefficients of the production functions by solving the system given by the first-order conditions of the two regions emissions at the Nash equilibrium, as well as their production functions. Thus our system is given by

$$\begin{aligned} a_i\,(b_i - e_i^u) &= \beta\, d_i\, \gamma^2\,(e_i^u + e_{-i}^u), && (17)\\ a_{-i}\,(b_{-i} - e_{-i}^u) &= \beta d_{-i}\,\gamma^2\,(e_i^u + e_{-i}^u), && (18)\\ \mathrm{GDP}_i &= a_i\, e_i^u\,(b_i - e_i^u/2), && (19)\\ \mathrm{GDP}_{-i} &= a_{-i}\, e_{-i}^u\,(b_{-i} - e_{-i}^u/2). && (20) \end{aligned}$$

The variables $\mathrm{GDP}_i$ and $\mathrm{GDP}_{-i}$ are regional GDP levels in 2010. Given that we obtain that all four parameters $a_i, a_{-i}, b_i, b_{-i}$ are positive, we accept that this is as a reasonable calibration for our model.

The regional split-up that we are working with is based on the World Bank regional aggregates and combines Sub-Saharan Africa, Middle East & North Africa, and Latin America & Caribbean into region $-i$, and the rest of the world in region $i$. The countries in this region $-i$ tend to be the ones that are impacted the strongest by climate change, and thus individually tend to have high incentives to undertake SRM. The damage parameter $\psi_i$ is taken at the median level for each region; results with the average level are very similar. GDP is measured at constant 2010 trillion USD, while the emissions are the carbon emissions measured in gigatons carbon. For the 2030 GDP data we use the SSP1 scenario which is consistent with the carbon path of the A2 SRES scenario. The data is given in Table 1. This gives us the following estimates for the parameters of the production

Table 1: Regional data

| Variable | Year | Data |
|---|---|---|
| $e_i^u$ | 2010 | 7.33 GtC |
| $e_{-i}^u$ | 2010 | 1.28 GtC |
| $\text{GDP}_i$ | 2010 | 56.00 trillion USD (constant 2010) |
| $\text{GDP}_{-i}$ | 2010 | 9.39 trillion USD (constant 2010) |
| $\text{GDP}_i$ | 2030 | 101.75 trillion USD (constant 2010) |
| $\text{GDP}_{-i}$ | 2030 | 26.26 trillion USD (constant 2010) |
| $\psi_i$ | | 1.6 % GDP/3°C warming |
| $\psi_{-i}$ | | 7.1 % GDP/3°C warming |

function

$$a_i = 2.078 \tag{21a}$$

$$a_{-i} = 11.329 \tag{21b}$$

$$b_i = 7.341 \tag{21c}$$

$$b_{-i} = 1.288 \tag{21d}$$

We then calculate $\bar{e}_{-i}$ and $\bar{e}$, which is given by

$$\bar{e}_{-i} = c_1/((p_{-i}\, d_{-i} + (1 - p_{-i})\bar{d}_{-i})\gamma\beta),$$

which yields $\bar{e} = 8.568$ GtC. Our calibration yields also an $\bar{e}_i = 9.512$ GtC. In addition, aggregate emissions at the UNE, or basically the current level of world emissions, is equal to $e^u = 8.617$ GtC.

### 3.2.2 Results

Now we can solve for the US and the CS. The conditions describing these two types of equilibria for the heterogeneous payoff functions are developed in Appendix D.

Clearly, changing $p_i$ or $\bar{d}_i$ has the same implications on the model outcome, so in fact

they boil down to one degree of freedom and not two.[8] In our sensitivity analysis we will thus simply change $\bar{d}_i$, knowing that this is equivalent to changing $p_i$, and only provide the numbers for $\tilde{d}_i$. Our objective is to show that, even in this very simple model and in the worst case scenario where the whole burden of the CS equilibrium lies with region 1, there exist reasonable parameters that yield us the result that region 1 prefers to play CS, while region 2 is better off in the CS than in the US equilibrium. These results are presented in Table 2.

As the results in Table 2 show, at least in the vicinity of our calibration but also for a large range of the $c_2$ parameter, we find that the equilibrium in conditional strategies is the sub-game perfect equilibrium. This equilibrium is still the extreme case where there is no burden sharing and region $i$ bears the whole cost of emission reductions. One could very well imagine that, in the case of burden sharing with region $-i$ also investing in emission reductions, there are larger ranges of parameters for which the equilibrium in conditional strategies is the sub-game perfect equilibrium.

Table 2: Numerical results

| $\Delta_i$ | $\Delta_{-i}$ | $c_2$ | $\Pi_i^r$ | $\Pi_i^n$ | $\Pi_{-i}^r$ | $\Pi_{-i}^n$ |
|---|---|---|---|---|---|---|
| 0.408525 | 0.413159 | .1 | 55.8988 | 55.8921 | 9.28744 | 9.28624 |
| 0.408525 | 0.413159 | 100 | 55.8988 | 55.8914 | 9.28744 | 9.28624 |
| 0.408525 | 0.413159 | .001 | 55.8988 | 55.8926 | 9.28744 | 9.28624 |
| 0.395586 | 0.413159 | .1 | 55.8988 | 55.8954 | 9.28744 | 9.28624 |

# 4 Conclusion

In this article we investigate whether a unilateral threat to employ SRM can induce players to commit to strategies with increased mitigation effort compared to what they would do in the absence of commitment.

[8] By this we mean that for any change in $\bar{d}_i > d_i$ we can find a change in $p_i$ that yields us the same expected damage as before.

For that purpose, we develop a two-region, (two-stage and) two-period game where regions choose emissions (or mitigation) and SRM. We first analyze a general model and discuss the potential outcomes, in terms of equilibrium and strategy. In the first period, regions choose their greenhouse gas emission levels, which yields a private benefit, but causes future climate damages, which are heterogeneous across world regions. In the second period, depending on aggregate emissions, regions can decide to undertake solar radiation management (SRM) at some cost. SRM is a means to reduce temperature, but at the same time it is perceived as a risky activity since it potentially involves a series of negative environmental impacts. This is captured by assuming that positive SRM induces a shift from a certain damage to a (larger) expected damage. Solving for the Nash equilibrium in SRM strategies allows us to identify a critical emission threshold that triggers (unilateral) SRM by the region that is the most vulnerable to climate change.

The main part of the analysis then consists in investigating how the potential deployment of SRM in the future affects current emissions. In the first period, regions have two options. Either they commit to meeting the threshold and face the resulting coupled constraint on aggregate emissions that prevents for SRM deployment in the second period. Or they act non-cooperatively and choose their emissions knowing that this will trigger (unilateral) SRM. Each scenario provides us with a candidate for the subgame perfect equilibrium (SPE).

We ultimately want to know under which conditions the threat of SRM may act as a coordination device by inducing regions to play the former SPE. This boils down to comparing the payoffs obtained at the two SPEs, which requires to study a more specific linear-quadratic application. We find that under some conditions involving the cost and benefit of SRM deployment, both regions may find it individually optimal to refrain from emitting too much in order to avoid SRM. This conclusion holds true even in the worst scenario in which one region – the one that will not undertake SRM – bears alone the responsibility of meeting (or not) the constraint. Finally, we calibrate the linear-quadratic model to real-world data and show that indeed there exist plausible ranges for the parameters under which commitment to prevent from future SRM use Pareto

dominates the SPE characterized by too high emissions and the use of SRM.

Our main take-away message is then somewhat different from the literature that adopts a more centralized perspective and emphasizes that the option to undertake SRM in the future should induce regions to reduce their current mitigation efforts (Jamieson, 1996; Keith, 2000; Quaas et al., 2017). We find that if regions can coordinate, then they may very well chose a level of emissions that is so low (or a mitigation level that is sufficiently high) such that SRM in the second period will not be worthwhile. The difference in results can be explained by the role of strategic interactions and the interpretation of SRM as a coordination device.

In terms of future research, we suggest that it would be useful to extend our result, namely that there exist coordination mechanisms which turn a free driver into a team player, to more general settings. In particular, future research may approach these questions: What other coordination strategies exist that take incentives away to free drive? What interrelation is there between free drivers and free riders? Can we design coordination strategies that bring both free riders and free drivers together?

# Appendix

# A Proof of Proposition 1

## A.1 Unconstrained Nash equilibrium (UNE)

Aggregate emissions $e^u$ and the two regions' emissions are:

$$e^u = \frac{2\,a\,b}{a+d_i+d_{-i}}$$
$$e^u_j = \frac{1}{a}\left(a\,b - d_j\,e^u\right)$$

Under condition (8), the UNE is well-defined – i.e. $e^u_j \geq 0$ for all $j$ – iff

$$d_{-i}\,e^u \leq a\,b \quad \Leftrightarrow \quad a \geq d_{-i} - d_i > 0. \tag{A.22}$$

This requires the difference between damage parameters is bounded (limited heterogeneity w.r.t to climate damage).

The two regions' payoffs at the UNE are given by (multiplied by $2a$):

$$\Pi^u_j = (ab)^2 - d_j(a+d_j)(e^u)^2 \text{ for } j = i, -i. \tag{A.23}$$

## A.2 Second period game

With the linear-quadratic specification, the condition $C'(0) = \mathbb{E}\left[D'_{-i}(\bar{e})\right]$ for the threshold level $\bar{e}$ becomes $c_1 = d_{-i}\,\bar{e}$. Thus, $e^u > \bar{e}$ if and only if (9) holds.

## A.3 First period problem

### A.3.1 US equilibrium

The amount of SRM is determined by $d_{-i}\ (e^n - g_{-i}) = c_1 + c_2\,g_{-i}$, and thus

$$g_{-i}(e) = \frac{d_{-i}\,e^n - c_1}{d_{-i} + c_2} \tag{A.24}$$

Equations (5) characterizing the US equilibrium simplify to

$$a\ (b - e_i^n) = \frac{c_2}{c_2 + d_{-i}}\ \tilde{d}_i\ \frac{c_1 + c_2\, e^n}{c_2 + d_{-i}} \tag{A.25a}$$

$$a\ \left(b - e_{-i}^n\right) = d_{-i}\ \frac{c_1 + c_2\, e^n}{c_2 + d_{-i}} \tag{A.25b}$$

Summing the two equations, we have

$$2\, a\, b - a\, e^n = \left(\frac{c_2}{c_2 + d_{-i}}\ \tilde{d}_i + d_{-i}\right)\ \frac{c_1 + c_2\, e^n}{c_2 + d_{-i}} \tag{A.26}$$

$$e^n = \frac{2\, a\, b\ (c_2 + d_{-i})^2 - \left(c_2\ \left(\tilde{d}_i + d_{-i}\right) + d_{-i}^2\right)\ c_1}{a\ (c_2 + d_{-i})^2 + \left(c_2\ \left(\tilde{d}_i + d_{-i}\right) + d_{-i}^2\right)\ c_2} \tag{A.27}$$

Thus,

$$\begin{aligned} e_i^n &= b - \frac{c_2}{c_2 + d_{-i}}\ \frac{\tilde{d}_i}{a}\ \frac{c_1 + c_2\, e^n}{c_2 + d_{-i}} \\ e_{-i}^n &= b - \frac{d_{-i}}{a}\ \frac{c_1 + c_2\, e^n}{c_2 + d_{-i}} \end{aligned} \tag{A.28}$$

The US equilibrium is well-defined iff $e_j^n \geq 0$ for all $j$ and $e^n > \bar{e}$. The latter inequality is equivalent to

$$\frac{2\, a\, b\ (c_2 + d_{-i})^2 - \left(c_2\ \left(\tilde{d}_i + d_{-i}\right) + d_{-i}^2\right)\ c_1}{a\ (c_2 + d_{-i})^2 + \left(c_2\ \left(\tilde{d}_i + d_{-i}\right) + d_{-i}^2\right)\ c_2} > \frac{c_1}{d_{-i}} \tag{A.29}$$

Rearranging leads to (12).

In addition, under the conditions used so far, it is straightforward to check that individual emissions are non-negative without any further restriction provided that $ab > c_1$: the marginal benefit from the first unit of emission is larger than marginal cost from the first unit of SRM.

Finally, regions' payoffs at the US equilibrium (multiplied by $2a$):

$$\Pi_i^n = (ab)^2 - \tilde{d}_i \left( a + \left( \frac{c_2}{c_2 + d_{-i}} \right)^2 \tilde{d}_i \right) (\tilde{e}^n)^2 , \tag{A.30a}$$

$$\Pi_{-i}^n = (ab)^2 - d_{-i} \, (a + d_{-i}) \, (\tilde{e}^n)^2 - d_{-i} \, a \, (\tilde{e}^n)^2 - c_1 \tag{A.30b}$$

$$= (ab)^2 - d_{-i} \, (a + d_{-i}) \, (\tilde{e}^n)^2 + \frac{a d_{-i}^2}{c_2} \left( \bar{e}^2 - (\tilde{e}^n)^2 \right) , \tag{A.30c}$$

where $\tilde{e}^n$ is the engineered temperature

$$\tilde{e}^n \equiv e^n - q_{-i}(e^n) = \frac{c_1 + c_2 \, e^n}{c_2 + d_{-i}}$$

### A.3.2 CS equilibrium

Solving the problem faced by the regions when they commit to not exceeding the threshold, we obtain the conditions for the coupled constraint Nash equilibrium, which is unique for a given $\varepsilon$:

$$e^r(\varepsilon) = \bar{e}, \tag{A.31a}$$

$$e_i^r(\varepsilon) = \frac{1}{a} \left[ a \, b \, (1 - 2\varepsilon) + (\varepsilon \, (a + d_{-i}) - (1 - \varepsilon) \, d_i) \, \bar{e} \right] , \tag{A.31b}$$

$$e_{-i}^r(\varepsilon) = \frac{1}{a} \left[ -a \, b \, (1 - 2\varepsilon) + ((1 - \varepsilon) \, (a + d_i) - \varepsilon \, d_{-i}) \, \bar{e} \right] . \tag{A.31c}$$

At this equilibrium, we observe that $e_i^{r\prime}(\varepsilon) < 0$ and $e_{-i}^{r\prime}(\varepsilon) > 0$. Moreover, if we want this solution to be well-defined for any $\varepsilon$, one must impose: $e_j^r(\varepsilon) \geq 0$ for all $j$, for all $\varepsilon$, which is equivalent to

$$\bar{e} \geq \max_j \left\{ \frac{a \, b}{a + d_j} \right\} \quad \Leftrightarrow \quad a \, b \, d_{-i} - c_1 \, (a + d_i) \leq 0. \tag{A.32}$$

This basically requires that a single region cannot bind the constraint alone when the other region bears the entire burden. This is somehow very demanding but needed to avoid unnecessary complications (the alternative is to define the interval of variation of $\varepsilon$ compatible with non-negative emissions).

The two regions' payoffs at the SPE with commitment in the first period are, multiplied by $2a$,

$$\Pi_i^r(\varepsilon) = (1-4\varepsilon^2)(ab)^2 + 4ab(\varepsilon(a+d_{-i}) - (1-\varepsilon)d_i)\varepsilon\bar{e} - (ad_i + (\varepsilon(a+d_{-i}) - (1-\varepsilon)d_i)^2)\bar{e}^2, \tag{A.33a}$$

$$\Pi_{-i}^r(\varepsilon) = -(1-2\varepsilon)(3-2\varepsilon)(ab)^2 + 4ab((1-\varepsilon)(a+d_i) - \varepsilon d_{-i})(1-\varepsilon)\bar{e} - (ad_{-i} + ((1-\varepsilon)(a+d_i) - \varepsilon d_{-i})^2)\bar{e}^2. \tag{A.33b}$$

# B Proof of Proposition 2

While $\Pi_i^r(1)$ does not depend on $\tilde{d}_i$, as there is no SRM in the CS equilibrium, $\Pi_i^n$ is decreasing in $\tilde{d}_i$. Let us denote the difference $\Pi_i^r(1) - \Pi_i^n$ as the function $G(\tilde{d}_i)$, which is increasing in $\tilde{d}_i \in [\underline{\Delta}_i, \overline{\Delta}_i)$. A necessary condition for unilateral commitment by region $i$ is $\lim_{\tilde{d}_i \to \overline{\Delta}_i} G(\tilde{d}_i) > 0$, which is equivalent to:

$$4\,a\,b\,((a+d_{-i})\,\bar{e} - a\,b) + \left(\overline{\Delta}_i\left(a + \left(\frac{c_2}{c_2+d_{-i}}\right)^2\overline{\Delta}_i\right) - a\,d_i - (a+d_{-i})^2\right)\bar{e}^2 > 0$$

After straightforward computations, one obtains that this inequality reduces to $\overline{\Delta}_i > d_i$, which holds under (8). This in turn implies that there exists a unique $\hat{\Delta}_i < \overline{\Delta}_i$ such that $\Pi_i^n \leq \Pi_i^r(1)$ for all $\tilde{d}_i \geq \hat{\Delta}_i$.

# C Proof of Proposition 3

First, note that payoffs at the CS satisfy: $\Pi_i^{r\prime}(\varepsilon) < 0$, $\Pi_{-i}^{r\prime}(\varepsilon) > 0$ for all $\varepsilon \in [0,1]$ (cf. Appendix A.3.2). It is also easy to check that $\Pi_i^r(0) > \Pi_i^n$ and $\Pi_{-i}^r(1) > \Pi_{-i}^n$: Both regions would prefer the CS if the burden were to be fully imposed to the other region. In addition, if $\tilde{d}_i < \hat{\Delta}_i$ then we know that $\Pi_i^n > \Pi_i^r(1)$. Finally, region $-i$ would be better off in US than in CS when it would have to bear the full burden, $\Pi_{-i}^r(0) < \Pi_{-i}^n$. We thus

consider the situation where both regions would prefer the US to the CS where they bear all the constraint.

In this case, from the direct comparison of the two regions' payoffs in the US and the CS equilibrium (see Appendices A.3.1 and A.3.2), it is possible to determine a critical sharing rule for each region according to which they prefer the CS equilibrium. Indeed, the difference $\Pi_j^n - \Pi_j^r(\varepsilon)$ gives a second order polynomial in $\varepsilon$ for $j = i, -i$. Solving these polynomials, we obtain

$$\Pi_i^r(\varepsilon) \gtreqless \Pi_i^n \qquad \Leftrightarrow \qquad \varepsilon \lesseqgtr \overline{\varepsilon}_i = \frac{\sqrt{\Gamma}_i - d_i\bar{e}}{(a + d_i + d_{-i})(e^u - \bar{e})} \tag{C.34}$$

$$\Pi_{-i}^r(\varepsilon) \gtreqless \Pi_{-i}^n \qquad \Leftrightarrow \qquad \varepsilon \gtreqless \underline{\varepsilon}_{-i} = \frac{(a + d_i + d_{-i})(e^u - \bar{e}) + d_{-i}\bar{e} - \sqrt{\Gamma}_{-i}}{(a + d_i + d_{-i})(e^u - \bar{e})} \tag{C.35}$$

with

$$\Gamma_i = d_i^2\bar{e}^2 + 2\,a\,(\Pi_i^r(0) - \Pi_i^n) \qquad \text{and}$$
$$\Gamma_{-i} = [(a + d_i + d_{-i})(e^u - \bar{e}) + d_{-i}\bar{e}]^2 - 2\,a\,(\Pi_{-i}^n - \Pi_{-i}^r(0)).$$

Both $\overline{\varepsilon}_i$ and $\underline{\varepsilon}_{-i}$ are positive under the above conditions which guarantee $e^u > \bar{e}$, $\Pi_i^r(0) - \Pi_i^n > 0$, and $\Pi_{-i}^n - \Pi_{-i}^r(0) > 0$.

A sharing rule where both regions are better off under CS than under US exists if and only if $\underline{\varepsilon}_{-i} \lesseqgtr \overline{\varepsilon}_i$. Comparing these thresholds directly would be a tedious exercise. Rather, we situate them with respect to the uniform rule $\varepsilon^u = \frac{1}{2}$. We obtain:

$$\overline{\varepsilon}_i \gtreqless \frac{1}{2} \Leftrightarrow \qquad \Pi_i^r(0) - \Pi_i^n \gtreqless \frac{1}{8\,a}(a + d_i + d_{-i})(e^u - \bar{e})(4d_i\bar{e} + (a + d_i + d_{-i})(e^u - \bar{e})) \tag{C.36a}$$

$$\underline{\varepsilon}_{-i} \lesseqgtr \frac{1}{2} \Leftrightarrow \quad \Pi_{-i}^n - \Pi_{-i}^r(0) \lesseqgtr \frac{1}{8\,a}(a + d_i + d_{-i})(e^u - \bar{e})(4d_{-i}\bar{e} + 3(a + d_i + d_{-i})(e^u - \bar{e})) \tag{C.36b}$$

A rough reading of these conditions provides us with a straightforward conclusion: region $i$'s gain (when moving from the US to the CS with $\varepsilon = 0$) should be sizable whereas the loss of region $-i$ should not be too high. We now have to find the conditions under which the two inequalities above are satisfied.

For region $-i$ we have

$$2\,a\,\left(\Pi^r_{-i}(0)-\Pi^n_{-i}\right)=(a+d_i+d_{-i})(e^u-\bar{e})[(a+d_i+d_{-i})((e^u-\bar{e})+2d_{-i}\bar{e}] \\ +d_{-i}(a+d_{-i}+\frac{ad_{-i}}{c_2})(\bar{e}^2-\tilde{e}^2).$$

Using this, the first inequality in (C.36) can be rewritten as:

$$(a+d_i+d_{-i})(e^u-\bar{e})\left[\frac{1}{4}(a+d_i+d_{-i})((e^u-\bar{e})+d_{-i}\bar{e}\right]<-d_{-i}\left(a+d_{-i}+\frac{ad_{-i}}{c_2}\right)\left(\bar{e}^2-\tilde{e}^2\right).$$

Now, $\Pi^u_{-i}<\Pi^n_{-i}$ is equivalent to:

$$d_{-i}(a+d_{-i})\left((e^u)^2-\bar{e}^2\right)<-d_{-i}\left(a+d_{-i}+\frac{ad_{-i}}{c_2}\right)\left(\bar{e}^2-\tilde{e}^2\right).$$

So it is sufficient to impose

$$(a+d_i+d_{-i})(e^u-\bar{e})\left[\frac{1}{4}(a+d_i+d_{-i})((e^u-\bar{e})+d_{-i}\bar{e}\right]<d_{-i}(a+d_{-i})\left((e^u)^2-\bar{e}^2\right),$$

to get the result. This condition can be rewritten as:

$$\frac{\bar{e}}{e^u}>\frac{(a+d_i+d_{-i})^2-4d_{-i}(a+d_{-i})}{(a+d_i+d_{-i})^2-4d_id_{-i}}.$$

For region $i$, we can define $\Pi^n_{-i}-\Pi^r_{-i}(0)$ as a function of $\tilde{d}_i$, for $\tilde{d}_i$ varying in the interval $(\underline{\Delta}_i,\hat{\Delta}_i)$, where $\hat{\Delta}_i$ has been defined as the solution of $\Pi^r_i(1)=\Pi^n_i$. Denote this function by $g(.)$, with:

$$g(\tilde{d}_i)=\frac{\tilde{d}_i}{2\,a}\left(a+\left(\frac{c_2}{c_2+d_{-i}}\right)^2\tilde{d}_i\right)\tilde{e}^2-d_i(a+d_i)\bar{e}^2.$$

We have: $g(\tilde{d}_i)>0$, $g'(\tilde{d}_i)>0$ for all $\tilde{d}_i$ and,

$$g(\hat{\Delta}_i)=(a+d_i+d_{-i})(e^u-\bar{e})((a+d_i+d_{-i})(e^u-\bar{e})+2d_i\bar{e}).$$

A necessary and sufficient condition for the existence of some $\tilde{d}_i\in(\underline{\Delta}_i,\overline{\Delta}_i)$ satisfying the second inequality in (C.36) is

$$g(\hat{\Delta}_i)>\tfrac{1}{4}(a+d_i+d_{-i})(e^u-\bar{e})(4d_i\bar{e}+(a+d_i+d_{-i})(e^u-\bar{e})),$$
$$\Leftrightarrow 3(a+d_i+d_{-i})(e^u-\bar{e})+d_i\bar{e}>0$$

which holds.

We can finally conclude that there exists a unique $\tilde{\Delta}_i < \hat{\Delta}_i$ such that $\bar{\varepsilon}_i \geq \frac{1}{2}$ for all $\tilde{d}_i \geq \tilde{\Delta}_i$. Note that it might be that $\tilde{\Delta}_i < \underline{\Delta}_i$, in which case, the result holds without any restriction.

# D Equilibrium conditions with heterogeneous regions

Here we develop the equilibrium conditions with heterogeneous regions that we use in Section 3.2.2.

The unconditional scenario US is described by the following system of conditions:

$$
\begin{aligned}
a_i(b_i - e_i^n) &= \beta\gamma(p_i d_i + (1-p_i)\bar{d}_i)(\gamma(e_i^n + e_{-i}^n) - g_i^n - g_{-i}^n), && \text{(D.37)}\\
a_{-i}(b_{-i} - e_{-i}^n) &= \beta\gamma(p_{-i} d_{-i} + (1-p_{-i})\bar{d}_{-i})(\gamma(e_i^n + e_{-i}^n) - g_i^n - g_{-i}^n), && \text{(D.38)}\\
g_i^n &= \max\left\{0, \frac{\beta(p_i d_i + (1-p_i)\bar{d}_i)(\gamma(e_i^n + e_{-i}^n) - g_{-i}^n) - c_1}{\beta(p_i d_i + (1-p_i)\bar{d}_i) + c_2}\right\}, && \text{(D.39)}\\
g_{-i}^n &= \max\left\{0, \frac{\beta(p_{-i} d_{-i} + (1-p_{-i})\bar{d}_{-i})(\gamma(e_i^n + e_{-i}^n) - g_i^n) - c_1}{\beta(p_{-i} d_{-i} + (1-p_{-i})\bar{d}_{-i}) + c_2}\right\} && \text{(D.40)}
\end{aligned}
$$

Based on these solutions we calculate the indirect profits in the non-committed equilibrium

$$
\begin{aligned}
\Pi_i^n &= a_i e_i^n (b_i - \frac{e_i^n}{2}) - \beta\frac{(p_i d_i + (1-p_i)\bar{d}_i)}{2}(\gamma(e_i^n + e_{-i}^n) - g_{-i}^n)^2 \\
&\quad -c_1\, g_1^n - c_2/2(g_i^n)^2. && \text{(D.41)}\\
\Pi_{-i}^n &= a_{-i} e_{-i}^n (b_{-i} - \frac{e_{-i}^n}{2}) - \beta\frac{(p_{-i} d_{-i} + (1-p_{-i})\bar{d}_{-i})}{2}(\gamma(e_i^n + e_{-i}^n) - g_{-i}^n)^2 \\
&\quad -c_1 g_{-i}^n - c_2/2(g_{-i}^n)^2. && \text{(D.42)}
\end{aligned}
$$

We solve the commitment equilibrium CE, firstly for the case where region $i$ bears the full costs of adhering to the constraint. In this case we solve the following system

$$
\begin{aligned}
a_{-i}(b_{-i} - e_{-i}^r) &= \beta d_{-i}\gamma^2\bar{e}, && \text{(D.43)}\\
e_i^r &= \bar{e} - e_{-i}^r. && \text{(D.44)}
\end{aligned}
$$

Due to our extensions of discounting, $\beta$ and the need for an additional parameter $\gamma$ to calibrate the emissions-temperature relationship, we have two additional parameters here compared to the theory part. The $d_i$ parameter used in the theory section then corresponds to $\beta\gamma^2\tilde{d}_i$.

We then calculate the indirect profits in the committed equilibrium, which are given by

$$\Pi_i^r = a_i e_i^r (b_i - \frac{e_i^r}{2}) - \beta \frac{(p_i d_i + (1-p_i)\bar{d}_i)}{2} (\gamma(e_i^r + e_{-i}^r))^2, \tag{D.45}$$

$$\Pi_{-i}^r = a_{-i} e_{-i}^r (b_{-i} - \frac{e_{-i}^r}{2}) - \beta \frac{(p_{-i} d_{-i} + (1-p_{-i})\bar{d}_{-i})}{2} (\gamma(e_i^r + e_{-i}^r))^2. \tag{D.46}$$

We then compare indirect profits between the non-committed and committed equilibrium in order to know whether region 1 has an incentive to play the committed equilibrium. The condition is $\Pi_i^r > \Pi_i^n$. We firstly check whether SRM deployment is a dominant strategy for region 2 (conditions for $\Pi_{-i}^u < \Pi_{-i}^n$), and then we assess if region 1 has an incentive to endorse the responsibility of cutting emissions in order to prevent SRM deployment by region 2 (conditions for $\Pi_i^r > \Pi_i^n$).

In the second scenario we allow for coupled constraints and study whether regions want to share the burden imposed by not exceeding the threshold $\bar{e}$. We analyze which sharing rule makes both regions better off under commitment.

The coupled constraint equilibrium is solved as follows. For $\varepsilon \in [0,1]$, with $\varepsilon\lambda \equiv \lambda_i$ and $\varepsilon_{-i} \equiv \lambda_{-i}$, we get

$$a_i(b_i - e_i^r) = \beta\gamma^2 d_i (e_i^r + e_{-i}^r) + \varepsilon\lambda, \tag{D.47}$$

$$a_{-i}(b_{-i} - e_{-i}^r) = \beta\gamma^2 d_{-i} (e_i^r + e_{-i}^r) + (1-\varepsilon)\lambda, \tag{D.48}$$

$$e_i^r + e_{-i}^r = \bar{e}. \tag{D.49}$$

This yields indirect profits given by

$$\Pi_i^r = a_i e_i^r (b_i - \frac{e_i^r}{2}) - \beta \frac{d_i}{2} (\gamma(e_i^r + e_{-i}^r))^2, \tag{D.50}$$

$$\Pi_{-i}^r = a_{-i} e_{-i}^r (b_{-i} - \frac{e_{-i}^r}{2}) - \beta \frac{d_{-i}}{2} (\gamma(e_i^r + e_{-i}^r))^2. \tag{D.51}$$

# References


Allen, M. R., Ingram, W. J., 2002. Constraints on future changes in climate and the hydrologic cycle. Nature 419 (6903), 224–232.

Aswathy, V. N., Boucher, O., Quaas, M., Niemeier, U., Muri, H., Mülmenstädt, J., Quaas, J., 2015. Climate extremes in multi-model simulations of stratospheric aerosol and marine cloud brightening climate engineering. Atmospheric Chemistry and Physics 15 (16), 9593–9610.

Barrett, S., 2008. The incredible economics of geoengineering. Environmental and Resource Economics 39 (1), 45–54.

Crutzen, P. J., 2006. Albedo enhancement by stratospheric sulfur injections: A contribution to resolve a policy dilemma? Climatic Change 77 (3), 211–220.

Emmerling, J., Tavoni, M., 2018. Climate engineering and abatement: A 'flat' relationship under uncertainty. Environmental and Resource Economics 69 (2), 395–415.

Ferraro, A. J., Charlton-Perez, A. J., Highwood, E. J., 2014. A risk-based framework for assessing the effectiveness of stratospheric aerosol geoengineering. PLoS ONE 9 (2), e88849 EP –.

Gardiner, S., Jamieson, D., Caney, S. (Eds.), 2010. Climate Ethics: Essential Readings. Oxford University Press, Oxford.

Heutel, G., Moreno-Cruz, J., Shayegh, S., 2016. Climate tipping points and solar geoengineering. Journal of Economic Behavior & Organization 132, 19–45.

Heyen, D., Wiertz, T., Irvine, P. J., 2015. Regional disparities in srm impacts: the challenge of diverging preferences. Climatic Change 133 (4), 557–563.

Horton, J., 2011. Geoengineering and the myth of unilateralism: Pressures and prospects for international cooperation. Stanford Journal of Law, Science, and Policy 4, 56–69.

IPCC, 2013. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jamieson, D., 1996. Ethics and intentional climate change. Climatic Change 33 (3), 323–336.

Keith, D. W., 2000. Geoengineering the climate: History and prospect. Annual Review of Energy and the Environment 25 (1), 245–284.

Klepper, G., Rickels, W., 2014. Climate engineering: economics prospects and considerations. Review of Environmental Economics and Policy 8 (2), 270–289.

Krawczyk, J. B., 2005. Coupled constraint nash equilibria in environmental games. Resource and Energy Economics 27 (2), 157–181.

Latham, J., 1990. Control of global warming? Nature 347, 339–340.

Le Quéré, C., Andrew, R. M., Friedlingstein, P., Sitch, S., Pongratz, J., Manning, A. C., Korsbakken, J. I., Peters, G. P., Canadell, J. G., Jackson, R. B., Boden, T. A., Tans, P. P., Andrews, O. D., Arora, V. K., Bakker, D. C. E., Barbero, L., Becker, M., Betts, R. A., Bopp, L., Chevallier, F., Chini, L. P., Ciais, P., Cosca, C. E., Cross, J., Currie, K., Gasser, T., Harris, I., Hauck, J., Haverd, V., Houghton, R. A., Hunt, C. W., Hurtt, G., Ilyina, T., Jain, A. K., Kato, E., Kautz, M., Keeling, R. F., Klein Goldewijk, K., Körtzinger, A., Landschützer, P., Lefèvre, N., Lenton, A., Lienert, S., Lima, I., Lombardozzi, D., Metzl, N., Millero, F., Monteiro, P. M. S., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-I., Nojiri, Y., Padin, X. A., Peregon, A., Pfeil, B., Pierrot, D., Poulter, B., Rehder, G., Reimer, J., Rödenbeck, C., Schwinger, J., Séférian, R., Skjelvan, I., Stocker, B. D., Tian, H., Tilbrook, B., Tubiello, F. N., van der Laan-Luijkx, I. T., van der Werf, G. R., van Heuven, S., Viovy, N., Vuichard, N., Walker, A. P., Watson, A. J., Wiltshire, A. J., Zaehle, S., Zhu, D., 2018. Global carbon budget 2017. Earth System Science Data 10 (1), 405–448.

Merk, C., Pönitzsch, G., Kniebes, C., Rehdanz, K., Schmidt, U., 2015. Exploring public perceptions of stratospheric sulfate injection. Climatic Change 130 (2), 299–312.

Moreno-Cruz, J. B., 2015. Mitigation and the geoengineering threat. Resource and Energy Economics 41, 248–263.

Moreno-Cruz, J. B., Ricke, K. L., Keith, D. W., 2012. A simple model to account for regional inequalities in the effectiveness of solar radiation management. Climatic Change 110 (3), 649–668.

Moreno-Cruz, J. B., Smulders, S., 2017. Revisiting the economics of climate change: the role of geoengineering. Research in Economics 71 (2), 212–224.

Morgan, J., Prieur, F., 2013. Global emission ceiling versus international cap and trade: What is the most efficient system to solve the climate change issue? Environmental Modeling & Assessment 18 (5), 493–508.

Moriyama, R., Sugiyama, M., Kurosawa, A., Masuda, K., Tsuzuki, K., Ishimoto, Y., 2017. The cost of stratospheric climate engineering revisited. Mitigation and Adaptation Strategies for Global Change 22 (8), 1207–1228.

Quaas, J., Quaas, M. F., Boucher, O., Rickels, W., 2016. Regional climate engineering by radiation management: Prerequisites and prospects. Earth's Future 4 (12), 618–625.

Quaas, M. F., Quaas, J., Rickels, W., Boucher, O., 2017. Are there reasons against open-ended research into solar radiation management? A model of intergenerational decision-making under uncertainty. Journal of Environmental Economics and Management 84, 1 – 17.

Ricke, K. L., Morgan, M. G., Allen, M. R., 2010. Regional climate response to solar-radiation management. Nature Geoscience 3, 537–541.

Ricke, K. L., Rowlands, D. J., Ingram, W. J., Keith, D. W., Granger Morgan, M., 2012. Effectiveness of stratospheric solar-radiation management as a function of climate sensitivity. Nature Clim. Change 2 (2), 92–96.

Robock, A., 2008. 20 reasons why geoengineering may be a bad idea. Bulletin of the Atomic Scientists 64 (2), 14–18.

Rosen, J., 1965. Existence and uniqueness of equilibrium point for concave n-person games. Econometrica 33, 520–534.

Roson, R., Sartori, M., 2016. Estimation of climate change damage functions for 140 regions in the gtap9 database. World Bank Policy Research Working Paper 7728.

Schelling, T. C., 1996. The economic diplomacy of geoengineering. Climatic Change 33 (3), 303–307.

Tidball, M., Zaccour, G., 2009. A differential environmental game with coupling constraints. Optimal Control Applications and Methods 30 (2), 197–207.

Weitzman, M., 2009. On modeling and interpreting the economics of catastrophic climate change. Review of Economics and Statistics 91 (1), 1–19.

Weitzman, M., 2011. Fat-tailed uncertainty in the economics of catastrophic climate change. Review of Environmental Economics and Policy 5 (2), 275–292.

Weitzman, M. L., 2015. A voting architecture for the governance of free-driver externalities, with application to geoengineering. The Scandinavian Journal of Economics 117 (4), 1049–1068.