Catastrophe and cooperation

Pim Heijnen* Lammertjan Dam†

October 3, 2016

Abstract

We study international environmental agreements in a setting that incorporates catastrophic climate change, and sovereign countries, who are heterogeneous in their exposure to climate change. This leads to a stochastic game with an absorbing state whose equilibrium structure is very different from the infinitely repeated games that are usually studied in the literature on environmental agreements. In particular there is no "folk theorem" that guarantees that the social optimum can be sustained in a Nash equilibrium as long as players are sufficiently patient. However, in most circumstances, it is feasible to implement an abatement scheme with a level of aggregate abatement that is close to the social optimum. Moreover, the discount rate has a non-monotonic effect on the optimal environmental agreement.

JEL-codes: C63, C73, H41, Q54
Keywords: international environmental agreement, voluntary participation, abatement, catastrophes, non-cooperative game

*Corresponding author: Faculty of Economics and Business, University of Groningen, P.O. Box 800, 9700AV, Groningen, e-mail: p.heijnen@rug.nl. We would like to thank Florian Wagener, Priscilla Man and seminar participants at the University of Amsterdam, the University of Queensland, the Tinbergen Institute, the University of Groningen, the ISDG workshop (Barcelona, July 2013), the 13$^{\text{th}}$ workshop on optimal control, dynamic games and nonlinear dynamics (Vienna, May 2015), SING11-GTM2015 (St. Petersburg, July 2015), and the 22$^{\text{nd}}$ annual conference of the EAERE (Zürich, June 2016) for their helpful comments and remarks.

†Faculty of Economics and Business, University of Groningen, e-mail: l.dam@rug.nl

1 Introduction

The two main economic questions regarding climate change are (i) which policy measures should be taken to combat the negative effects of climate change and (ii) how do we design international environmental agreements to implement these policy measures? In this paper, our focus is on the latter question. We develop a parsimonious model of international environmental agreements. We argue that the three key issues that shape the form of international environmental agreements are that climate change is catastrophic, that countries are sovereign, and that countries differ in their exposure to climate change. In this setting, we characterize the optimal stable environmental agreement and show that it can be close to the social planner outcome.

By catastrophic we mean an abrupt change in the climate. For instance, the rise in global temperatures could trigger the melting of the Siberian permafrost. The subsequent release of methane would lead to a further increase in temperature, leading to the release of more methane and even further increase in temperature. This example is just one possible scenario, but catastrophic shifts in ecological systems are a well-documented phenomenon (Scheffer et al., 2001). Because catastrophes involve a great deal of uncertainty, both in when it will happen and what precisely will happen, the economic cost will be large compared to the cost due to any gradual change. For tractability, we focus on the case where the only cost of climate change is the cost of a catastrophic shift. Moreover, the catastrophe is a random event and the probability that the catastrophe occurs decreases if resources are allocated to abatement. The recent literature (see Polasky et al. (2011) and the references therein) has devoted much attention to this aspect of climate change, focusing on the optimal choice of one decision maker. However, it is the joint (or aggregate) level of abatement that determines by how much this probability decreases, and our contribution is to extend the analysis to multiple decision makers, i.e. countries.

Climate change is global in scale, so limiting the negative effects requires international cooperation. The Kyoto protocol shows that the world is aware of the necessity for coop-

eration. Unfortunately, as the failure of the USA to ratify the Kyoto protocol illustrates, it also shows that any international environmental agreement needs to entice countries to participate: given that all other countries join, it should be optimal for a country to join as well. This imposes constraints on the form an international environmental agreement can take (see Barrett (1994, 2003) for an analysis of the deterministic case).

Participation constraints will differ between countries, since some countries will be more severely affected by a climate catastrophe. For instance, a rise in sea level is a serious issue for a low-lying country like the Netherlands, whereas the direct cost for a country without coastal areas, like Switzerland, will be zero.

These three features are modeled in the following way.[1] There are two states of the world: pre-catastrophe and post-catastrophe. Pre-catastrophe, all countries have the same level of net production. The catastrophe permanently destroys a fraction of net production, where the fraction differs between countries. In each period of time a catastrophic shift happens with some probability. Countries can allocate resources to abatement: the higher aggregate abatement, the lower the probability of a catastrophe.

First, we compare the social optimum to the stationary Nash equilibrium. As expected, the Nash equilibrium is inefficient. The first source of inefficiency is that there is not enough abatement in the Nash equilibrium. The second source of inefficiency is more subtle. In general, in our framework, welfare decreases if prior to the catastrophe some countries abate more than others, i.e. given an aggregate level of abatement, welfare is highest when all countries abate the same amount. Since the incentive to abate is stronger if a country is hurt more by the catastrophe, in the Nash equilibrium the level of abatement will differ between countries and this is an additional cause for welfare to decrease.

Second, we examine stable international environmental agreements, i.e. an international

[1]Interestingly, Dutta and Radner (2006, 2009) claim to address the same three features in their model of international environmental agreements. However their model is deterministic and abatement enters both the objective function and the state equation in a linear fashion. While this allows them to fully characterize the set of Nash-equilibria even when countries are heterogeneous, none of the features of a catastrophic shift appear in their approach.

environmental agreement in which every country joins and cooperation is sustained by trigger strategies. Since the outside option for some countries is more attractive than for others, the distribution of abatement among countries tends to be unbalanced. This imbalance implies that in general the social optimum cannot be implemented by a stable international environmental agreement. However, in most circumstances, it is feasible to implement an abatement scheme with the same level of aggregate abatement as the social optimum. The difficulty is to persuade all countries to join this abatement scheme. Countries with little exposure to the negative effects of climate change will only join an international environmental agreement if their abatement requirements are low. The burden than falls disproportionally on countries that are severely impacted by the catastrophe. As discussed in the previous paragraph, welfare decreases if abatement is less evenly distributed among countries. Therefore, in the optimal stable international environmental agreement aggregate abatement will be (slightly) less than in the social optimum (but substantially higher than in the Nash equilibrium).

Third, most of the literature on international environmental agreements focuses on (infinitely) repeated games. Our model is a stochastic game with an absorbing state. In this setting the usual folk theorems do not apply and we show that very patient players may actually cause lower levels of abatement in the optimal stable international environmental agreement. Note that one critique of the Stern report (Stern, 2007) has been that it overemphasizes the cost of climate change by choosing a very low discount rate (Nordhaus (2007) is the most vocal critic). We provide one reason why this critique may not be valid: if abatement is mainly an instrument to prevent catastrophes, then its benefits are not long-run, but rather the benefits occur before the catastrophe takes place. This encourages a somewhat impatient decision maker to invest in abatement, but a very patient decision maker will disregard it.

Our paper brings together two strands of the literature.[2] There is an extensive literature

[2]Due to the inherent dynamics of the problem, we focus on the part of the literature, which the dynamics is explicit. Another approach is Barrett (2013), who models the climate catastrophe as a (static) threshold public good game, where passing the (potentially unknown) threshold is interpreted as a cli-

on the stability of international environmental cooperation (van der Ploeg & de Zeeuw, 1992; Fuentes-Albero & Rubio, 2010; Breton, Sbragia, & Zaccour, 2010). In this literature, there are usually immediate benefits of abatement, since abatement marginally improves the state of the environment. We focus on non-marginal improvements, since one of the benefit of abatement is that it might postpone (or even avoid) a catastrophe. We are not the first to investigate catastrophic shifts: see for instance Tahvonen and Salo (1996); Nævdal (2001); Mäler, Xepapedeas, and de Zeeuw (2003); Wagener (2003); Heijdra and Heijnen (2013). However, most of these papers focus on a single decision maker or, occasionally, multiple decision makers. But when this literature considers the case of multiple decision makers they do not focus on the question whether the cooperative outcome can be sustained in a Nash-equilibrium.[3] This paper is an attempt to investigate these issues in a simple framework.

The outline of the paper is as follows. Section 2 introduces the model. Theoretical results are presented in section 3. A numerical example is presented in section 4 and we discuss the role of the discount factor and country heterogeneity. In Section 5, the effect of irreversibility is analyzed. Section 6 concludes. All proofs are in the appendix.

2 Model

2.1 The environment

In order to get tractable results, the representation of the environmental catastrophe will be extremely parsimonious, i.e. we only distinguish between a pre-catastrophe state of the world and a post-catastrophe state of the world. While there are differences in environmental quality within each of these states, these are unimportant compared to the huge

mate catastrophe. While this captures the idea that improvements (or deteriorations) are non-marginal, it disregards the fact that it is more costly to reverse climate change.

[3] A good, recent example of this is van der Ploeg and de Zeeuw (2014): their modeling framework can be seen as a more general version of our model. However, they only compare the cooperative and the noncooperative outcome without addressing the question whether cooperation is stable.

deterioration of environmental quality as a result of the catastrophe. Moreover, the timing of the catastrophe is uncertain.

Formally, the state of the environment at time $t = 0, 1, 2, \ldots$ is denoted by Ω_t. The environment is either in a good state ($\Omega_t = G$) or the environment is in a bad state ($\Omega_t = B$). The good state is pre-catastrophe and the bad state is post-catastrophe. We start in the pre-catastrophe world: $\Omega_0 = G$. In each period there is a probability p of staying in the good state, the bad state is irreversible. [4]

2.2 The economy

There are n countries, indexed by $i = 1, \ldots n$. Each country internally follows the Golden Rule and maximizes net production. The catastrophe reduces net production because it reduces the marginal productivity of capital. Net production in country i is y (if $\Omega_t = G$) and $\alpha_i y$ (if $\Omega_t = B$), where $\alpha_i \in (0, 1)$.[5] The effect of a catastrophe is a decrease in net production and α_i is a measure of how much country i is hit by the catastrophe. Countries are labeled such that $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$, i.e. we rank countries from least to most hit. Country i invests $m_i \geq 0$ in abatement. The remainder of net production is consumed and gives country i an instantaneous utility of $u(y - m_i)$ (if $\Omega_t = G$) or $u(\alpha_i y - m_i)$ (if $\Omega_t = B$), where the utility function $u(\cdot)$ is increasing, strictly concave and satisfies the

[4] One example to motivate this reduced form approach is the Allee effect. The normal growth model for the population of a species is the logistic growth model, enriched with an Allee effect. The Allee effect is the phenomenon that if the population falls below a certain critical size, then the population is no longer sustainable and will go extinct. Suppose that the population is not extinct and at the steady state level. In addition, suppose that there are random events that influence the population size, such as droughts or a decreased presence of predators. As long as the impact of these random events is small, population size will quickly return to the steady state level. However, a severe negative shock could push the population level below the minimum threshold and cause extinction. In terms of our model, this is a catastrophic shift from G to B which happens with probability p, i.e. the probability of a sufficiently large negative shock.

[5] Net production is production minus investment in capital. For example, suppose that the production function is $f(k) = \sqrt{\alpha k}$, where $\alpha = 1$ pre-catastrophe and $\alpha = \alpha_i$ post-catastrophe. Moreover, let ψ denote the depreciation rate. Then net production is $f(k) - \psi k$. Under the golden rule, net production is maximized: $\max_k f(k) - \psi k = \frac{\alpha}{4\psi}$. Define $y = \frac{1}{4\psi}$ and we see that net production is y before the catastrophe and $\alpha_i y$ after.

Inada conditions. Moreover, the countries are prudent: $u''' \geq 0$.[6] Countries maximize the discounted sum of instantaneous utility, which is referred to as the welfare of country i. Welfare is normalized by multiplication with a factor $(1-\delta)$.

Let

$$Y_i(\Omega_t) = \begin{cases} y & \text{if } \Omega_t = G \\ \alpha_i y & \text{if } \Omega_t = B \end{cases}$$

denote net production and let m_{it} denote abatement of country i at time t. Then the welfare of country i at time t is

$$V_i(\Omega_t) = (1-\delta)\mathbb{E}\sum_{s=t}^{\infty} \delta^{t-s} u(Y_i(\Omega_t) - m_{it}),$$

where $\delta \in (0,1)$ is the discount factor. Welfare can be written recursively as

$$V_i(\Omega_t) = (1-\delta)u(Y_i(\Omega_t) - m_{it}) + \delta \mathbb{E}_t[V_i(\Omega_{t+1}) \mid m_{1t}, \cdots, m_{nt}]. \tag{1}$$

In principle, country i may choose a different level of abatement each period. However, our focus will be on stationary behavior, where abatement only depends on the state. Consequently, the time subscript is frequently dropped.

2.3 Abatement and welfare

Let $M = \sum_i m_i$ denote aggregate abatement. We assume that the transition probability depends on aggregate abatement: $p(M)$, where $p(\cdot)$ is increasing, concave and $p < 1$.

Remark that due to the irreversibility of the bad state, there will be no abatement post-catastrophe. Furthermore, in all cases we examine abatement is time-invariant. Therefore, m_i will denote the level of abatement of country i pre-catastrophe. An abatement scheme is a vector $(m_1, m_2, \ldots, m_n, M)$, where $M = \sum_i m_i$. Using the recursive formulation in (1),

[6] Given that in our model countries abate to minimize the probability of a catastrophe, we have this assumption in common with the literature on optimal loss prevention (Eeckhoudt & Gollier, 2005). However that literature deals with static loss prevention with a single decision maker whereas we study dynamic loss prevention with multiple decision makers.

welfare in the good state for country i in this abatement scheme is denoted by $V_i(m_i, M)$ and is implicitly given by

$$V_i(m_i, M) = (1-\delta)u(y-m_i) + \delta p(M)V_i(m_i, M) + \delta(1-p(M))u(\alpha_i y),$$

where the last term is the discounted welfare of being in the bad state. Hence:

$$V_i(m_i, M) = \frac{(1-\delta)u(y-m_i) + \delta(1-p(M))u(\alpha_i y)}{1-\delta p(M)}.$$

It can easily be shown that V_i is decreasing in m_i and increasing in M.

2.4 Comparison to other models

When formulated in this manner, we appear to have a standard public good game where the welfare of country i depends negatively on its own level of abatement and positively on the aggregate level of abatement. However, this is only true if all countries abate at a constant level when $\Omega_t = G$. The Nash equilibria we discuss in Section 3 are shaped primarily by dynamic consideration (what is the gain of deviating from a certain abatement scheme?). Even the social planner solution (which effectively maximizes the sum of the welfare of the countries) has some unusual features such as the fact that the benefit (which depend on M) and the cost of abatement (which depend on m_i) enter the welfare function non-separably. For instance, an increase in total abatement leads to a decrease in the marginal cost of abatement for each country. This is effect is usually absent in public good games.

To be able to highlight some of the differences between our approach and the approach as taken by Barrett (1994) and the subsequent literature, let us briefly sketch a model without catastrophic shifts. Suppose that aggregate abatement yields immediate benefits $B(M, \alpha)$, which is an increasing, concave function (in M) that satisfies the Inada-conditions. Countries are still heterogeneous: the benefit of country i is $B(M, \alpha_i)$, where $\frac{\partial B}{\partial \alpha_i} < 0$, and country 1 is still affected least by environmental problems, or phrased alternatively, receives the largest benefits. Then the welfare of each country in each period is:

$$W_i(m_i, M) = u(y-m_i) + B(M, \alpha_i)$$

The stage game is one where each country simultaneously sets abatement. The stage game is infinitely repeated. From the Folk Theorem we know that if for each country welfare in the socially optimal outcome exceeds welfare in the Nash equilibrium of the stage game and the discount factor is sufficiently close to 1, then the socially optimal outcome can be enforced.

As Dutta (1995) and Levine (2000) show, these results do not generalize to stochastic games with absorbing states.[7] In the game presented here, due to the absorbing state it becomes difficult to punish very patient players. Note that punishment is only possible in the good state (in the bad state utility is always equal to $u(\alpha_i y)$). Since very patient players put little weight on the present (good) state, it may not be possible to set punishments at an appropriately high level. Both in terms of interpretation and in terms of the structure of the set of equilibria, a stochastic game differs from the standard infinitely repeated game.

3 Theoretical results

In this section, we derive the equilibrium conditions for three different scenarios and characterize their properties. The benchmark is the social planner solution (SP), where abatement levels are chosen such that joint welfare is maximized. Then we examine a stationary Nash equilibrium (NE), where all countries chose abatement independently. Finally, we examine the joint welfare maximizing Nash-equilibrium that can be sustained using trigger strategies. We will refer to the final scenario as a stable international environmental agreement (SA).

[7]Note that if we fix the player's strategies in a stochastic game, then this induces a Markov chain over the state space. Dutta (1995) shows that if this Markov chain is irreducible for any choice of the player's strategies, then the set of equilibrium payoffs approaches the entire individually rational set of payoffs as the discount factor approaches one (which is a version of the Folk Theorem). Levine (2000) shows that this result does not hold when the Markov chain is reducible.

3.1 Social planner

The social planner solution is to maximize

$$\sum_i V_i(m_i, M)$$

subject to

$$\sum_i m_i = M \text{ and } m_i \geq 0 \text{ for all } i$$

Before we present the solution, note that if the social planner want to implement a desired level of aggregate abatement, then the efficient way to do achieve this is by setting the same level of abatement in each country. Basically, this is the equimarginal principle in action: the social planner allocates abatement to the country with the lowest marginal cost of abatement. In the optimum, the marginal cost of abatement needs to be the same. But since the countries are identical before the catastrophe, this means that abatement is the same in all countries. Let $m_1 = m_2 = \ldots = m_n \equiv \mu$. Define $M^{SP} = n\mu$ as the level of aggregate abatement in the social planner solution. We make the following assumption:

Assumption 1. *There is an interior social planner solution:* $\mu > 0$.

This is obviously the interesting case to examine: the divergence between the socially optimal outcome and the competitive outcome arises because usually in the latter case there is underabatement. This situation only occurs if the social planner abates a strictly positive amount.

Note that if W is the maximum aggregate welfare (i.e. welfare in the social planner solution), then by the principle of optimality:

$$W = \max_{\mu} n(1-\delta)u(y-\mu) + \delta p(n\mu)W + \delta(1-p(n\mu))\sum_i u(\alpha_i y)$$

Hence, the optimal level of abatement in each country is determined by

$$-n(1-\delta)u'(y-\mu) + n\delta p'(n\mu)[W - \sum_i u(\alpha_i y)] = 0. \tag{2}$$

Substituting $W = \sum_i V_i(\mu, n\mu)$ yields, after some tedious algebra, an implicit expression for abatement per country in the social planner solution:

$$\left[nu(y-\mu) - \sum_i u(\alpha_i y) \right] f(n\mu) = u'(y-\mu), \tag{3}$$

where $f(M) \equiv \delta p'(M)/(1-\delta p(M))$.[8]

3.2 Stationary Nash equilibrium

In the Nash-equilibrium, in each period every country independently sets its abatement level. Note that this is a stochastic game with a finite state space. A common equilibrium concept is a stationary equilibrium, where the strategy does not depend on history or time. In our setting, this means that we have to determine the level of abatement for each country when the environment is in the good state. Note that an abatement scheme $(m_1, m_2, \ldots, m_n, M)$ will yield welfare $V_i(m_i, M)$ to country i. Then we apply the one-stage deviation principle to find the equilibrium level: for each country, it should not be welfare-improving to deviate from m_i at any single stage of the game.[9] Formally:

Definition 1. *An abatement scheme* $(m_1^{NE}, m_2^{NE}, \ldots, m_n^{NE}, M^{NE})$ *is a stationary Nash equilibrium when, for all* i*,*

$$m_i^{NE} \in \arg\max_{m \geq 0} (1-\delta)u(y-m) + \delta p(M_{-i}^{NE} + m) V_i^{NE} + \delta(1 - p(M_{-i}^{NE} + m))u(\alpha_i y),$$

where $V_i^{NE} = V_i(m_i^{NE}, M^{NE})$ *and* $M_{-i}^{NE} = \sum_{j \neq i} m_j^{NE}$.

Using the definition, we see that m_i^{NE} is determined by

$$-(1-\delta)u'(y - m_i^{NE}) + \delta p'(M^{NE}) \left[V_i^{NE} - u(\alpha_i y) \right] \leq 0, \tag{4}$$

where the inequality holds if $m_i^{NE} > 0$.[10] Substituting

$$V_i^{NE} = \frac{(1-\delta)u(y - m_i^{NE}) + \delta(1 - p(M^{NE}))u(\alpha_i y)}{1 - \delta p(M^{NE})},$$

[8] Note that due to concavity of u and p, the first-order condition in (2) is a necessary and sufficient for a maximizer. In the appendix, we show that (3) has a unique solution.

[9] Observe that although each country selects a single abatement level, this level is determined by dynamic considerations.

[10] Note that due to concavity of u and p, the first-order condition in (4) is a necessary and sufficient for a maximizer.

we get

$$-(1-\delta)u'(y-m_i^{NE})+\delta p'(M^{NE})\left[\frac{(1-\delta)u(y-m_i^{NE})+\delta(1-p(M^{NE}))u(\alpha_i y)}{1-\delta p(M^{NE})}-u(\alpha_i y)\right]\leq 0$$

which simplifies to

$$\left[u(y-m_i^{NE})-u(\alpha_i y)\right]\frac{\delta p'(M^{NE})}{1-\delta p(M^{NE})}\leq u'(y-m_i^{NE}),$$

or

$$\left[u(y-m_i^{NE})-u(\alpha_i y)\right]f(M^{NE})\leq u'(y-m_i^{NE}) \text{ for all } i. \tag{5}$$

Compared to (3) we see that in the Nash equilibrium country i only takes into account its own benefit of abatement (i.e. $\left[u(y-m_i^{NE})-u(\alpha_i y)\right]$). This is the classic freeriding problem.

In general it is hard to proof existence of equilibria in stochastic games, especially when the action space is not finite or countable infinite.[11] Therefore, we present the following condition under which the Nash-equilibrium does not only exist, but is also unique.

Proposition 1. *If f is decreasing, then there is a unique stationary Nash-equilibrium.*

Note that when f is decreasing, abatement is a strategic substitute and, therefore, a country will abate less if other countries abate more. Formally:

Assumption 2. *Abatement is a strategic substitute, i.e. f is decreasing.*

The stationary Nash-equilibrium has the following properties:

Proposition 2. *Abatement is weakly increasing: $0\leq m_1^{NE}\leq\cdots\leq m_n^{NE}$. Moreover:*

1. *Suppose $j<k$. Then $m_k^{NE}=0$ implies $m_j^{NE}=0$.*
2. *$m_k^{NE}=m_j^{NE}>0$ if and only if $\alpha_k=\alpha_j$.*

Countries, that are more severely affected by the catastrophe, will abatement more. Moreover, it is possible that the least affected countries do not abate at all.

[11] Stochastic games were introduced by Shapley (1953), who shows that Markov perfect equilibria exist when the action and state space are finite. Although similar theorems now exist for more general games, existence theorems do not exist for the most general case. See the discussion in Fudenberg and Tirole (1991, pp. 503–505).

Proposition 3. *In the stationary Nash equilibrium the aggregate level of abatement is less than in the social planner solution:* $M^{NE} < M^{SP}$.

This shows that the Nash equilibrium is inefficient in two ways. There is not enough abatement and the abatement is not distributed efficiently among countries.

3.3 Stable international environmental agreements

In an international environmental agreement, the countries jointly agree on an abatement scheme. The agreement is supported by trigger strategies, i.e. if any country deviates from the agreement, then from that period onward we enter a punishment regime. For the moment, we will assume that if country i deviates, then it will be punished in such a manner that its welfare after deviation is at most the welfare it would receive in the stationary Nash-equilibrium i.e. $\tilde{V}_i \leq V_i^{NE}$. In the next section, where we present a numerical example, different punishment regimes are discussed.

The incentive constraints have two peculiar features. First, incentive constraints are not independent: overabatement by one country changes the incentives for the other countries. In particular, it makes it more attractive for other countries to deviate. Therefore, if one country voluntarily abates more, other countries may deviate from the optimal scheme. It is tempting to argue that if a country wants to abate more, then welfare can be increased by letting this country abate more and reducing the levels for the other countries. In general this is not true, since a greater spread in the abatement levels will decrease joint welfare. Hence, any deviation from the abatement scheme, including upward deviations, need to be punished. Second, it is not necessarily true that more patient players have less strict incentive constraints (for reasons outlined at the end of Section 2.4). Therefore, we expect that there is an optimal discount rate that is most conducive to cooperation.

An abatement scheme $(m_1, \cdots, m_n, M)$ leads to an incentive constraint for each country. If country i does not deviate, then it receives welfare $V_i(m_i, M)$. The most attractive deviation

gives welfare:

$$\max_{m\geq 0}(1-\delta)u(y-m)+\delta p(M_{-i}+m)\tilde{V}_i+\delta(1-p(M_{-i}+m))u(\alpha_i y).$$

Then the incentive constraint for country i is

$$IC_i:\ V_i(m_i,M)\geq \max_{m\geq 0}(1-\delta)u(y-m)+\delta p(M_{-i}+m)\tilde{V}_i+\delta(1-p(M_{-i}+m))u(\alpha_i y).$$

Observe that the RHS of the incentive constraint is a function of M_{-i}, i.e. abatement by all countries except i. It turns out that in the analysis, it is convenient to first investigate if a certain level of aggregate abatement can be sustained by an international environmental agreement. We say that an international environmental agreement is stable if the incentive constraint for all countries is satisfied. Conditional on M, we have $M_{-i}=M-m_i$ and both the LHS and the RHS of IC_i are functions of m_i and M.

Since it is trivial to enforce an abatement scheme in which $M=M^{NE}$, and since welfare can be increased by abating more, we focus on abatement schemes where $M>M^{NE}$. We can show the following.

Lemma 1. *Conditional on the aggregate level of abatement M, there exist a bound on abatement z_i such that country i will join an environmental agreement when its contribution m_i does not exceed z_i, i.e. $IC_i \implies 0\leq m_i\leq z_i(M)$.*

Therefore, countries maximize joint welfare under the following constraints:

$$\max_{m_i,M}\sum_i V_i(m_i,M)$$

such that

$$\sum_i m_i = M$$
$$0\leq m_i\leq z_i(M)\quad \text{for all } i$$

Conditional on the aggregate level of abatement, we can characterize how the burden will be shared among the countries. Note that conditional on M, maximizing $\sum_i V_i$ is equivalent to maximizing $\sum_i u(y-m_i)$. To find the optimal allocation, we make use of the following lemma:

Lemma 2. *Let $\tilde{m}$ be a feasible vector of abatement levels. Suppose that for some j and k, $0 \leq \tilde{m}_j < \tilde{m}_k \leq z_k$. Let $\hat{m} = \tilde{m} + \varepsilon\nu$, where ν is a vector such that $\nu_j = 1$, $\nu_k = -1$ and all remaining entries are zero. Then there exists $\varepsilon > 0$ such that $\hat{m}$ will strictly improve welfare and $\hat{m}$ is feasible.*

The lemma implies the following

1. All countries for which the upper bound is not binding ($m_i < z_i$) have the same level of abatement.

2. If for country i the upper boundary is binding ($m_i = z_i$), then this level of abatement is smaller than the level of abatement for the countries for which the upper bound is not binding.

Roughly speaking, in an optimal international environmental agreement the burden will be shared equally. However, the requirement that the agreement is stable may lead to deviations from this principle. In particular, countries with a binding incentive constraint are allowed to abate less to ensure that they will not deviate from the environmental agreement. Formally stated:

Proposition 4. *Suppose $M^{NE} < M < \sum_i z_i$. The solution to the maximization problem*

$$\max_{m_i} V_i(m_i, M)$$

such that

$$\sum_i m_i = M$$

and

$$0 \leq m_i \leq z_i \quad \textit{for all } i$$

is unique and can be determined as follows. Construct the function

$$\mathcal{F}(\gamma) = \gamma \sum_{i \notin U(\gamma)} 1 + \sum_{i \in U(\gamma)} z_i - M,$$

where

$$U(\gamma) = \{i \mid z_i \leq \gamma\}.$$

There exists a unique $\gamma^* > 0$ such that $\mathcal{F}(\gamma^*) = 0$. The solution is given by

$$m_i = z_i \quad \textit{for all } i \in U(\gamma^*),$$
$$m_i = \gamma^* \quad \textit{otherwise.}$$

It is possible for the social planner solution and the optimal stable agreement to coincide. Observe that if

$$\mu \leq \min_i z_i(M^{SP}), \tag{6}$$

then the social optimum is a stable agreement and must therefore be the optimal agreement. When (6) does not hold, stable agreements can still reach the same level of aggregate abatement as the social optimum. This is feasible if

$$M^{SP} \leq \sum_i z_i(M^{SP}). \tag{7}$$

However this may not be the optimal agreement, since in general it will require some countries to abate less than other countries. Ceteris paribus, a greater divergence of abatement among countries leads to a loss in welfare (in the sense of Lemma 2). By lowering the aggregate level of abatement, abatement per country can be more homogeneous. This leads to a tradeoff between the optimal amount of aggregate abatement and the efficient implementation of such a scheme. We expect that at the socially optimal level of abatement, the latter effect dominates the first, as the numerical results in the next section confirm.[12]

4 Numerical example

In this section, we discuss how the three different scenarios behave with the aid of a numerical example.[13] The parameter values and functional forms are as follows. For the

[12]We have assumed that the utility function is strictly concave. Most of our results hold when the utility function is linear with the notable exception of Lemma 2. With linear utility, the social welfare function only depends on aggregate abatement, i.e. the distribution of abatement is not of importance. Social welfare has a unique maximum at $M = M^{SP}$ and in the optimal stable agreement, aggregate abatement is as close to M^{SP} as the incentive constraints allow. Then (7) is the condition under which the social optimum and the optimal stable agreement coincide.

[13]Matlab-code for all computations are available on request.

transition probability, we use:

$$p(m) = \frac{\tau m + \varphi}{\tau m + 1},$$

where $\tau = 100$ and $\varphi = 0.1$. Note that without abatement the probability of staying in the good state is φ. The utility function is $u(c) = \sqrt{c}$. Moreover $y = 1$ and $\delta = 0.8$. We set $n = 5$ and $\alpha_i = 0.95 - 0.0375(i-1)$. Country 1 loses 5% of net production due to the catastrophe and country 5 loses 20%. We consider two punishment regimes. In the Nash-punishment scenario after deviation countries will play the stationary Nash-equilibrium. In the maxmin-punishment, all countries (except the deviator) will stop abatement completely. These two punishment regimes represent the two extremes with maxmin is the harshest punishment that the countries can inflict upon a deviator, while Nash-punishment is the most lenient one (without actually rewarding deviators).

The aggregate level of abatement in the social planner solution is 0.1116, and hence the abatement level per country is 0.0223. See Table 1 for the incentive constraints at this level of aggregate abatement. We see that under Nash-punishment the social planner solution is not feasible, since country 1, 2 and 3's maximum abatement level is below 0.0223. However, since $\sum_i z_i = 0.1326$, it is feasible to have the same level of aggregate abatement in the SEA. Under maxmin-punishment, the social planner solution is feasible. This shows that if punishment is severe enough then the social planner solution can be enforced by a stable environmental agreement. To see what shape the optimal agreement takes, we now focus our attention on the Nash-punishment scenario.

Table 2 shows the abatement level for each country in each different scenario, as well as aggregate abatement and the probability of staying in the good state. Though it is feasible to have the same level of aggregate abatement as in the SP, it is optimal to abate a bit less in the SA. In this case, the incentive constraint for the first four countries is binding and country 5 provides the remainder of the abatement. In the NE, abatement is considerable lower with the first three countries not abating at all. In both the SP and the SA, the probability of staying in the good state is approx. 92.6%. In the NE, this figure is a bit lower at 85.1%. While this may seem a relatively small difference, it implies that on average

it takes 13.5 periods to transition to the bad state in both the SP and the SA, but it only takes 6.7 periods in the NE.[14]

Table 3 shows the welfare for each country. Strikingly, in the SP countries 1 and 2 receive lower welfare than in the NE (which of course is compensated by the huge welfare gain of country 5). This is the reason why (even for small discount rates) the social planner solution cannot be enforced by a trigger strategy. Hence, country heterogeneity is an obstruction to reaching the first-best outcome.

In the previous section, we argued that there may be an optimal discount rate that is most conducive to cooperation. When the discount factor is 0.8 and maxmin-punishments are used, the social planner solution is a stable environmental agreement. The hypothesis is then that this ceases to be true when the discount factor is sufficiently close to one. In the numerical example, this happens at $\delta = 0.99996$. Hence it not true that if the social planner solution is a stable environmental agreement for a discount rate $\bar{\delta}$, then it is also stable for all discount rates $\delta > \bar{\delta}$. In that sense the effect of the discount rate on the stability of the social planner solution is non-monotonic.

5 Reversible catastrophes

Finally, we briefly consider reversible catastrophes to investigate to what extent the irreversibility matters for the result. We achieve this by introducing a probability ν that the state moves from B to G, independent of the amount of abatement in the bad state. This exercise should be regarded as nothing more than a sensitivity analysis. It is not meant to

[14]Note that the aggregate level of abatement is severely restricted by country 1, 2 and 3, whose willingness to contribute is much lower than country 4 and 5. Potentially, a partial coalition of country 4 and 5 could perform better than the "grand coalition" since it faces less strict incentive constraints. However, in the example, a partial coalition where country 4 and 5 cooperate performs worse than the grand coalition. Calculation show that if country 4 and 5 cooperate, then in the welfare-maximizing outcome (subject to the incentive constraint) the levels of abatement of country 4 and 5 are resp. 0.0216 and 0.0552 (and the associated welfare levels are 0.9675 and 0.9493). The welfare of the participating countries is lower than when all countries cooperate. Moreover, because aggregate abatement is also substantially lower, the welfare of the non-participating countries also decreases.

	z	
Country	Nash	Maxmin
1	0.0072	0.0236
2	0.0129	0.0365
3	0.0189	0.0476
4	0.0297	0.0579
5	0.0639	0.0686

Table 1: Incentive constraints when aggregate abatement is at the social planner level. The column "Nash" is Nash-punishment and the column "Maxmin" is the maxmin-punishment.

Country	SP	SA	NE
1	0.022	0.007	0.000
2	0.022	0.013	0.000
3	0.022	0.019	0.000
4	0.022	0.030	0.006
5	0.022	0.042	0.044
Aggregate	0.112	0.111	0.051
p	0.926	0.926	0.851

Table 2: Abatement in three different scenarios. In the stable environmental agreement Nash-punishments are used.

Country	SP	SA	NE
1	0.9856	0.9914	0.9906
2	0.9811	0.9848	0.9833
3	0.9766	0.9779	0.9759
4	0.9720	0.9690	0.9664
5	0.9672	0.9592	0.9476

Table 3: Welfare in three different scenarios. In the stable environmental agreement Nash-punishments are used.

	$\nu = 0.01$	$\nu = 0.05$
m_i	0.0539	0.0393
M	0.2696	0.1963
$\Pr[G]$	0.2370	0.5341
V_1	0.9742	0.9776
$\bar{V}_1$	0.9752	0.9769

Table 4: Reversible catastrophes, outcomes for different values of ν: m_i is abatement per country, M is aggregate abatement, $\Pr[G]$ is probability of being in the good state in the long run, V_1 is optimal average long-run welfare of country 1, $\bar{V}_1$ is the maxmin average long-run welfare of country 1.

imply that an environmental catastrophe will just resolve without deliberate human action.

Since both B and G are now recurring states, we can use the folk theorem from Dutta (1995) and therefore we focus on the case where $\delta = 1$.[15] Any abatement scheme results in a long-run distribution over the two states. If countries do not discount future payoffs, then they should maximize the long-run average payoff. Dutta (1995) shows that any feasible payoff that is individually rational (i.e. it yields a higher payoff then the maxmin-payoff) can be sustained by a Nash-equilibrium.

Table 4 shows the numerical results for the case where $\nu = 0.01$ and the case where $\nu = 0.05$. When $\nu = 0.01$, the catastrophe is more severe in the sense that after moving from G to B on average the bad state lasts for 100 periods compared to 20 periods when $\nu = 0.05$. This results in higher abatement in the good state when $\nu = 0.01$. Recall that country 1 has the least incentive to invest in abatement: we only need to check if the individually rational constraint holds for country 1. It turns out that the individual rationality constraint holds for country 1 when $\nu = 0.05$ but not when $\nu = 0.01$.[16] Only when the probability that the catastrophe is reversible is sufficiently high, the social welfare maximizing outcome can be sustained by a stable environmental agreement.

[15] All other parameters remain at the same value.

[16] The indifference point is at $\nu \approx 0.03386$.

6 Concluding remarks

In this paper, we develop a parsimonious model of international environmental agreements, incorporating three key issues: climate change is catastrophic, countries are sovereign (and hence there are participation constraints in designing international environmental agreement) and countries differ in their exposure to climate change. Technically, this leads to a stochastic game with an absorbing state whose equilibrium structure is very different from the infinitely repeated games that are usually studied in the literature on environmental agreements. Due to the irreversibility of the catastrophe, our intuition on discounting does not work. Since the catastrophe is irreversible, the payoff of a very patient player will be mainly determined by the payoff in the bad state. This limits the extent to which a player can be punished when it deviates from an abatement scheme. Hence, international environmental agreements could actually be easier to implement if decision makers are a bit myopic. If catastrophes are reversible, then "folk theorems" again apply and the main obstacle to implementing the social planner solution is the heterogeneity of countries: in this case side payments may be essential to foster international cooperation.

References

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

Barrett, S. (2003). *Environment and statecraft: the strategy of environmental treaty-making.* Oxford University Press.

Barrett, S. (2013). Climate treaties and approaching catastrophes. *Journal of Environmental Economics and Management*, *66*, 235–250.

Breton, M., Sbragia, L., & Zaccour, G. (2010). A dynamic model for international environmental agreements. *Environmental and Resource Economics*, *45*, 25–48.

Dutta, P. (1995). A folk theorem for stochastic games. *Journal of Economic Theory*, *66*, 1–32.

Dutta, P., & Radner, R. (2006). Population growth and technological change in a global warming model. *Economic Theory*, *29*, 251–270.

Dutta, P., & Radner, R. (2009). A strategic analysis of global warming: Theory and some numbers. *Journal of Economic Behavior and Organization*, *71*, 187–209.

Eeckhoudt, L., & Gollier, C. (2005). The impact of prudence on optimal prevention. *Economic Theory*, *26*, 989–994.

Fudenberg, D., & Tirole, J. (1991). *Game theory*. The MIT Press.

Fuentes-Albero, C., & Rubio, S. (2010). Can international environmental cooperation be bought? *European Journal of Operational Research*, *202*, 255–264.

Heijdra, B., & Heijnen, P. (2013). Environmental abatement and the macroeconomy in the presence of ecological thresholds. *Environmental and Resource Economics*, *55*, 47–70.

Levine, D. (2000). The castle on the hill. *Review of Economic Dynamics*, *3*, 330–337.

Mäler, K., Xepapedeas, A., & de Zeeuw, A. (2003). The economics of shallow lakes. *Environmental and Resource Economics*, *26*, 603–624.

Nævdal, E. (2001). Optimal regulation of eutrophying lakes, fjords, and rivers in the presence of threshold effects. *American Journal of Agricultural Economics*, *83*, 972–984.

Nordhaus, W. (2007). A review of the "Stern review on the economics of climate change". *Journal of Economic Literature*, 686–702.

Polasky, S., de Zeeuw, A., & Wagener, F. (2011). Optimal management with potential regime shifts. *Journal of Environmental Economics and Management*, *62*, 229–240.

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., & Walker, B. (2001, October). Catastrophic shifts in ecosystems. *Nature*, *413*, 591-596.

Shapley, L. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, *39*, 1095–1100.

Stern, N. (Ed.). (2007). *The economics of climate change: the Stern review*. Cambridge University press.

Tahvonen, O., & Salo, S. (1996). Nonconvexities in optimal pollution accumulation. *Journal of Environmental Economics and Management*, *31*, 160–177.

van der Ploeg, F., & de Zeeuw, A. (1992). International aspects of pollution control. *Environmental and Resource Economics*, *2*, 117–139.

van der Ploeg, F., & de Zeeuw, A. (2014). *Non-cooperative and cooperative responses to climate catastrophes in the global economy: A north-south perspective* (Tech. Rep.). University of Oxford.

Wagener, F. (2003). Skiba points and heteroclinic bifurcations, with applications to the shallow lake system. *Journal of Economic Dynamics and Control*, *27*, 1533–1561.

Appendix: Proofs

Proof that the social planner solution is unique We have to show that (3) has a unique solution. Consider

$$\left[nu(y-m) - \sum_i u(\alpha_i y) \right] f(nm) = u'(y-m)$$

as a function of m. Observe that the LHS and the RHS of the equation are continuous and differentiable functions in m. Note that the RHS is increasing in m. We show that evaluated at any solution the LHS is decreasing in m. Since by assumption 1 a solution exists, this implies uniqueness. The LHS is decreasing in m if

$$\left[nu(y-m) - \sum_i u(\alpha_i y)\right] f'(nm) < u'(y-m) f(nm). \tag{8}$$

From (3) we see that for any solution:

$$\left[nu(y-m) - \sum_i u(\alpha_i y)\right] = \frac{u'(y-\mu)}{f(n\mu)} \tag{9}$$

Evaluating (8) at $m = \mu$, substituting (9) and simplifying, we get:

$$f'(n\mu) < (f(n\mu))^2$$

Using the definition of f, this simplifies to $\delta p''(1-\delta p') < 0$ which is true.

Proof of Proposition 1 Suppose that the aggregate level of abatement is M. If this is the aggregate abatement of a stationary Nash-equilibrium, then either m_i is the solution to

$$\left[u(y-m_i) - u(\alpha_i y)\right] f(M) = u'(y-m_i).$$

or $m_i = 0$ when this solution does not exist. This defines a continuous function $m_i = \zeta_i(M)$. If f is decreasing, then it is straightforward to verify that there exists $\bar{M}_i$ such that $\zeta_i(M) = 0$ for all $M \geq \bar{M}_i$, and ζ_i is decreasing in $[0, \bar{M}_i]$.[17] Observe that in any stationary Nash-equilibrium $\sum_i \zeta_i(M) = M$. We show that $g(M) \equiv M - \sum_i \zeta_i(M)$ has a unique non-negative root. Note that g is continuous, $g(0) < 0$ and $g(\max_i \bar{M}_i) > 0$, where the last claim follows from the bound on ζ_i. Then by the intermediate value theorem, there g has a non-negative root. Moreover, g is increasing and therefore the root is unique.

[17] We assume that $\max_i \bar{M}_i > 0$. Note that if $\max_i \bar{M}_i = 0$, then trivially there is a unique Nash-equilibrium in which no country abates.

Proof of Proposition 2 To prove the first statement, note that $m_k^{NE} = 0$ implies that (5) reduces to

$$[u(y) - u(\alpha_k y)]\, f(M^{NE}) \leq u'(y)$$

Since $\alpha_j \geq \alpha_k$, we have

$$[u(y) - u(\alpha_j y)]\, f(M^{NE}) \leq [u(y) - u(\alpha_k y)]\, f(M^{NE}).$$

Hence

$$[u(y) - u(\alpha_j y)]\, f(M^{NE}) \leq u'(y)$$

and $m_j^{NE} = 0$.

To prove the second statement, note that abatement is positive and therefore the inequality in (5) holds:

$$\left[u(y - m_i^{NE}) - u(\alpha_i y)\right] f(M^{NE}) = u'(y - m_i^{NE}).$$

Given, that $u(\cdot)$ is strictly increasing, it is obvious that $m_k^{NE} = m_j^{NE} > 0$ if and only if $\alpha_j = \alpha_k$.

We prove the main claim by contradiction. Take two countries i and j such that $i < j$ (and therefore $\alpha_i > \alpha_j$) and suppose that $m_i^{NE} > m_j^{NE}$. Because of the first statement, we can focus on interior solutions without loss of generality. From (5), we get:

$$f(M^{NE}) = \frac{u'(y - m_i^{NE})}{u(y - m_i^{NE}) - u(\alpha_i y)} = \frac{u'(y - m_j^{NE})}{u(y - m_j^{NE}) - u(\alpha_j y)}.$$

Since $m_i^{NE} > m_j^{NE}$ and $u(\cdot)$ is concave, $u'(y - m_i^{NE}) > u'(y - m_j^{NE})$. This implies:

$$u(y - m_i^{NE}) - u(\alpha_i y) > u(y - m_j^{NE}) - u(\alpha_j y)$$
$$u(y - m_i^{NE}) - u(y - m_j^{NE}) > u(\alpha_i y) - u(\alpha_j y)$$

Note that the LHS of this inequality is negative and the RHS is positive. This contraction establishes that $m_i^{NE} \leq m_j^{NE}$.

Proof of Proposition 3 First, suppose that every country abates a strictly positive amount, i.e. for all i, equation (4) holds with equality. Then summing (4) over i, we get

$$\delta p'(M^{NE}) \left[\sum_i V_i^{NE} - u(\alpha_i y) \right] = (1-\delta) \sum_i u'(y - m_i^{NE}) \tag{10}$$

Let $\hat{m} = M^{NE}/n$ and let $\hat{V}_i = V_i(\hat{m}, M^{NE})$. Observe that $\sum_i \hat{V}_i > \sum_i V_i^{NE}$ since aggregate welfare increases when abatement is distributed more equally (for a given level of aggregate abatement) and, since $u''' \geq 0$, $\sum_i u'(y - m_i^{NE}) \geq n u'(y - \hat{m})$ by Jensen's inequality. From these observation and (10), we have

$$\delta p'(n\hat{m}) \left[\sum_i \hat{V}_i - u(\alpha_i y) \right] \geq (1-\delta) n u'(y - \hat{m}). \tag{11}$$

In the social planner solution, we have

$$\delta p'(n\mu) \left[W - u(\alpha_i y) \right] = (1-\delta) n u'(y - \mu). \tag{12}$$

Now suppose, contrary to the claim of the Proposition, that $\bar{m} \geq \mu$. Then

$$(1-\delta) n u'(y - \hat{m}) \geq (1-\delta) n u'(y - \mu) = \delta p'(n\mu) \left[W - u(\alpha_i y) \right],$$

where the equality follows from (12). Comparing this equation to (11), it must be that

$$\delta p'(n\hat{m}) \left[\sum_i \hat{V}_i - u(\alpha_i y) \right] \geq \delta p'(n\mu) \left[W - u(\alpha_i y) \right]$$

Note that due to concavity of p, we have $p'(n\hat{m}) \leq p'(n\mu)$. Therefore $\sum_i \hat{V}_i \geq W$, which contradicts the fact that W is defined as the (strict) maximum of total welfare. Hence $\mu > \bar{m}$ and $M^{SP} > M^{NE}$.

Second, we examine boundary equilibria. Suppose that $m_1^{NE} = 0, \ldots, m_k^{NE} = 0$ and $m_{k+1}^{NE} > 0, \ldots, m_n^{NE} > 0$. If the social planner would only take into account the welfare of country $k+1$ up to n, then the aggregate level of abatement would be more than the aggregate level of abatement in the stationary Nash equilibrium. When it also takes into account the welfare of country 1 up to k, the social planner will increase the aggregate level of abatement. Hence, $M^{NE} < n\mu$ *a fortiori*.

Proof of Lemma 1 Both the LHS and the RHS of IC_i are decreasing in m_i. The claim follows if we can show that the derivative of the LHS is strictly less than the derivative of the RHS. Suppose that in an abatement scheme country i has to abate m_i and aggregate abatement is M. Let m^* denote the optimal deviation from the abatement scheme. First we show that $m^* \leq m_i$.

Let m^* denote country i's optimal deviation. The aim is to show that $m^* \leq m_i$. Since country i's welfare from deviation is concave in m (cf. RHS of IC_i), it is sufficient to show that the derivate of welfare evaluated at m_i is negative:

$$\delta p'(M)[\tilde{V}_i - u(\alpha_i y)] \leq (1-\delta)u'(y - m_i)$$

Let $\sigma \equiv (u')^{-1}$. Therefore the inequality can be rewritten as:

$$m_i \geq y - \sigma\left(\frac{\delta p'(M)[\tilde{V}_i - u(\alpha_i y)}{1-\delta}\right),$$

since σ is decreasing. In general, we need a minimal level of m_i to guarantee that the optimal deviation is downward. Unless

$$0 \geq y - \sigma\left(\frac{\delta p'(M)[\tilde{V}_i - u(\alpha_i y)]}{1-\delta}\right),$$

or equivalently

$$\delta p'(M)[\tilde{V}_i - u(\alpha_i y)] \leq (1-\delta)u'(y).$$

Observe that

$$\delta p'(M)[\tilde{V}_i - u(\alpha_i y)] \leq \delta p'(M^{NE})[V_i^{NE} - u(\alpha_i y)] \leq (1-\delta)u'(y - m_i^{NE}) \leq (1-\delta)u'(y),$$

where the first inequality follows from $M > M^{NE}$, concavity of p and the fact that $\tilde{V}_i < V^{NE}$, the second inequality from the definition of the stationary Nash-equilibrium, and the final inequality from the concavity of u. Hence all deviations are downward: $m^* \leq m_i$.

Then from the first-order condition, we have:

$$-(1-\delta)u'(y - m^*) + \delta\left[\tilde{V}_i - u(\alpha_i y)\right]p'(M_{-i} + m^*) \leq 0.$$

Consequently:

$$0 < \delta \left[\tilde{V}_i - u(\alpha_i y)\right] p'(M_{-i} + m^*) \leq (1-\delta)u'(y - m^*). \tag{13}$$

Remark that the derivative of the LHS of IC_i to m_i is

$$\frac{-(1-\delta)u'(y-m_i)}{1-\delta p(M)} < 0$$

and the derivative of the RHS of IC_i to m_i is

$$-\delta \left[\tilde{V}_i - u(\alpha_i y)\right] p'(M_{-i} + m^*) < 0$$

Using (13), we see that it suffices to show that

$$\frac{-(1-\delta)u'(y-m_i)}{1-\delta p(M)} < -(1-\delta)u'(y-m^*) \leq -\delta \left[\tilde{V}_i - u(\alpha_i y)\right] p'(M_{-i} + m^*) < 0$$

The only unproven inequality is

$$\frac{-(1-\delta)u'(y-m_i)}{1-\delta p(M)} < -(1-\delta)u'(y-m^*)$$

which follows directly from the fact that $1 - \delta p(M) < 1$, concavity of the utility function and $m^* \leq m_i$.

Proof of Lemma 2 It is obvious that $\hat{m}$ is feasible for ε small enough. We have to show that:

$$\sum_i u(y - \tilde{m}_i) < \sum_i u(y - \hat{m}_i).$$

This is equivalent to showing that

$$u(y - \tilde{m}_j) + u(y - \tilde{m}_k) < u(y - \tilde{m}_j - \varepsilon) + u(y - \tilde{m}_k + \varepsilon)$$

Then using Taylor expansions, we get

$$u(y - \tilde{m}_j) + u(y - \tilde{m}_k) < u(y - \tilde{m}_j) - u'(y - \tilde{m}_j)\varepsilon + u(y - \tilde{m}_k) + u'(y - \tilde{m}_k)\varepsilon - \kappa_\varepsilon \varepsilon^2$$

for some $\kappa_\varepsilon \geq 0$ (since $u(\cdot)$ is concave). Therefore:

$$\kappa_\varepsilon \varepsilon < u'(y - \tilde{m}_k) - u'(y - \tilde{m}_j),$$

where the RHS is strictly positive by the strict concavity of $u(\cdot)$ and the assumption that $\tilde{m}_j < \tilde{m}_k$. Since $\lim_{\varepsilon \downarrow 0} \kappa_\varepsilon \varepsilon = 0$, there exists $\varepsilon > 0$ such that the inequality will hold.

Proof of Proposition 4 Suppose γ is the proposed level of abatement for each country whose incentive constraints are satisfies if they abate at this level and the aggregate level of abatement is M. Let $U(\gamma)$ be the set of countries for which the upper boundary is binding at this level of abatement:

$$U(\gamma) = \{i \mid z_i \leq \gamma\}.$$

Note that U is a strict subset of $\{1, \cdots, n\}$ since $M < \sum_i z_i$ by assumption. Then, by Lemma 2, the abatement scheme proposed in the proposition is a welfare-maximizing outcome if

$$\gamma \sum_{i \notin U(\gamma)} 1 + \sum_{i \in U(\gamma)} z_i = M.$$

Define

$$\mathcal{F}(\gamma) = \gamma \sum_{i \notin U(\gamma)} 1 + \sum_{i \in U(\gamma)} z_i - M$$

Note that $\mathcal{F}(0) = -M < 0$, $\mathcal{F}(\max_i z_i) = \sum_i z_i - M > 0$ and $\mathcal{F}$ is increasing since U is a strict subset of $\{1, \cdots, n\}$. By the intermediate value theorem, we have that there is a unique value of $\gamma \in (0, \max_i z_i)$ such that $\mathcal{F}(\gamma) = 0$.