A Dynamic Model for International Environmental Agreements

Abstract

In this paper we develop a model to analyze, in a dynamic framework, how countries join international environmental agreements (IEAs). In the model, where countries suffer from the same environmental damage as a result of total global emissions, a non-signatory country decides on its emission level by maximizing its own welfare, whereas a signatory country decides on its emission level by maximizing the aggregate welfare of all signatory countries. Signatory countries are assumed to be able to punish non-signatories at a cost. When countries decide on their pollution emissions, they account for the evolution of the stock of pollution over time. Moreover, we propose a mechanism to describe how countries reach a stable IEA. The model is able to capture situations characterized by partial cooperation within an IEA that is stable over time. It also captures situations where all countries participate in a stable agreement, and situations where no stable agreement is feasible. When more than one possibility coexist, the long-term outcome of the game depends on the initial conditions (i.e., the size of the initial group of signatory countries and the pollution level).

Key Words: Environment; international agreements; dynamic game; replicator dynamics; non-cooperative game.

1 Introduction

Environmental problems often share the feature of being international, that is, the welfare of one country depends not only on its own policy but also on those of other countries. Examples of these problems include acid rain, pollution emissions, and harvesting of fishing grounds or rain forests. In the case of pollution emissions, air pollutants combine in the atmosphere and increased concentrations of these gases depend on total, global emissions. Countries have realized that these kinds of problems have to be solved on a global basis and that international environmental agreements (IEAs) are the only solutions.

The participation of countries in an international agreement to improve the quality of the environment is a complex question for two main reasons. First, countries are sovereign and their participation in international agreements is voluntary. There is no supra-national authority that forces countries to participate in an agreement, nor is there an international environmental judicial system powerful enough to guarantee compliance to an IEA (Endres 2004; Wagner 2001; Finus 2000). Second, each country may have an incentive to free ride. In fact, while the costs for reducing emissions are carried exclusively by the country that is taking action, the benefits of a reduction in emissions are shared by all countries, so that each country has an incentive to wait for the others to reduce their emissions. This problem is commonly known as "The Tragedy of the Commons" (Hardin 1968). Because of these interactions, global environmental problems can be modeled in a game theoretic framework. The literature based on this approach has developed following two streams: the cooperative and non-cooperative approaches. In this paper, we adopt the non-cooperative point of view; the main concept, in this case, is that players cannot make binding agreements and they act as rivals and in their own best interest,[1] so that successful agreements must be *self-enforcing* (d'Aspremont et al. 1983). The conditions for an agreement to be self-enforcing are profitability, which ensures that accession to the agreement is individually rational for a country, and stability, which ensures that the group of signatory countries is an equilibrium.

The main body of literature on non-cooperative games and self-enforcing agreements uses a static framework to describe the problem of pollution emissions, that is, the environmental damage is assumed to depend on the flow of emissions. In this literature, following d'Aspremont et al. (1983), two conditions define an equilibrium: internal stability, which implies that no member has an incentive to leave the agreement, and external stability, which implies that no non-member has an incentive to join the agreement. Behind these concepts, there is the idea that governments can re-optimize their choice; however, in static models, this is only hypothetical, because, since the game will not be played again, there is no adjust-

[1] Cooperative game theory assumes that players realize that, if they act as a group and coordinate their actions, they can obtain mutual benefits. A particular agreement is the result of questions about the circumstances under which the agreement can be established, what can be achieved as a group, and the ways in which the benefits of the cooperation are redistributed among the participants (Dockner et al. 2000; Wagner 2001)

ment towards a stable solution and no change of state. In this framework, several stages in which specific decisions have to be made can be considered; a typical example is a two-stage game, where countries decide whether or not to become members of an IEA in the first stage (membership game), and decide on their emissions in the second stage (emissions game).

Many static models using this stability concept have reached the conclusion that successful cooperation among a large number of countries is difficult to achieve, and that the size of the membership of a stable IEA is inversely related to the relative extent of the environmental damage. In order to explain the level of participation that is observed in reality, like for example, in the Montreal Protocol,[2] static models have incorporated several ideas to increase the likelihood of cooperation. Some of them are: Stackelberg leadership of signatories (Barrett 1994; Diamantoudi and Sartzetakis 2006; Rubio and Ulph 2006), transfers (Carraro and Siniscalco 1993; Hoel and Schneider 1997), reputation effects (Hoel and Schneider 1997; Jeppensen and Andersen 1998; Cabon-Dhersin and Ramani 2006), issue linkages (Botteon and Carraro 1998; Le Breton and Soubeyran 1997; Barrett 1997; Katsoulacos 1997; Carraro and Siniscalco 1997, 1998; Mohr and Thomas 1998), including a minimum participation clause (Carraro et al. 2003), and considering modest emission-reduction targets (Finus 2004).

Although any model abstracts from reality, static games applied to pollution emissions and IEAs may be criticized on two important aspects. One aspect is the stock externality: transboundary environmental damage is usually related to the accumulation (stock) of pollution, rather than to the emissions (flow). The second aspect is that countries can revise their decisions to be or not to be member of an agreement at different points in time, especially if the environmental damage changes over time.

In the literature considering the dynamics of the pollution stock, the concepts of internal and external stability conditions may be conceptually linked to renegotiation-proof or dynamically consistent agreements, in the context of repeated and dynamic games, respectively. Part of this literature develops differential games that compare different solution concepts without addressing the membership problem (see for instance Long 1992; van der Ploeg and de Zeeuw 1992; Dockner and Long 1993; Dockner and Nishimura 1999, and Rubio and Casino 2002). Papers considering the membership decision include Rubio and Casino (2005); Germain et al. (2003), and Rubio and Ulph (2007). In Rubio and Casino (2005), the membership game is played once, for good, and on the basis of this result, signatory and non-signatory countries decide on their emissions by solving an infinite-horizon differential game in both open-loop and feedback strategies. Thus, the dynamics of the pollution stock affects the emissions strategies but not the IEA membership. In Germain et al. (2003), a transfer scheme is proposed in a dynamic game, such that full cooperation is maintained over time. In this case, the dynamics of the stock pollutant is included in the membership decision, but the transfer scheme is able

[2]The Montreal Protocol on Substances that Deplete the Ozone Layer, signed in 1987, currently includes 191 nations.

to ensure full cooperation so that the size of the IEA membership does not change over time. Finally in Rubio and Ulph (2007) the emergence and stability of IEAs over time is studied in a difference game with a dynamic pollution stock. In each period, countries solve both an emissions game and a membership game so that the number of signatories changes over time with the stock of pollution. The authors assume an upper bound on non-signatory emissions. They also suppose that, in each period, signatories are randomly selected, so that the stability concept implies the equality of the expected welfare of signatories and non-signatories. The result is the existence of a unique steady-state of pollution stock to which corresponds a steady-state for the size of a stable IEA. As in the static case, the size of the self-enforcing IEA is negatively related to the significance of the damage costs.

In this paper, we recognize that countries can join or quit an agreement over time and that it takes time to reach a stable IEA, and we propose a possible mechanism to describe how this could happen. We choose a discrete time setting because, even if pollution evolves continuously over time, countries' decisions regarding whether to enter or leave an IEA are rather made at discrete moments.

The agreement we model in this paper includes two clauses. The first, common in the literature (see, e.g., Barrett 1994; Rubio and Casino 2005), is that a signatory country agrees to decide on its emission level by maximizing the aggregate welfare of the coalition. On the other hand, a non-signatory country decides on its emission level by maximizing its individual welfare. In both cases, the countries solve a dynamic emissions game, that is, they use the total discounted welfare over an infinite horizon, taking into account the evolution of the pollution stock.

The second clause requires a signatory country to punish non-signatory countries: in becoming a member of the agreement, a country agrees to inflict a cost to non-members. In particular, we assume that the punishment is a function of the level of accumulated pollution. This is motivated by the assumption that, as the level of accumulated pollution increases, environmental concern of signatory countries also increases, so that the punishment they agree to inflict becomes stronger.

Including some sort of punishment in an IEA as a tool to increase and sustain cooperation is not uncommon in static games (Barrett 2003, 1997). This non-environmental cost suffered by non-signatories can be interpreted as some form of sanction related to international trade, or as a social norm. Indeed, among the measures used by the Montreal Protocol to boost participation in the agreement, there are trade sanctions against non-members (banning trade between signatories and non-signatories). We first consider an exogenous punishment, as in Hoel and Schneider (1997), then we compute the "optimal" punishment as the one that makes the long-run full cooperation solution stable to unilateral deviations.

Moreover, since a general punishment or, more specifically, a restriction of trade is not without adverse consequences, we also include a cost suffered by signatory countries for pun-

ishing the non-signatories. We assume this cost to be proportional to the punishment. This is an aspect that has not been addressed in Hoel and Schneider (1997), nor in general, to the best of our knowledge, in the context of IEAs to reduce pollution emissions.[3] Our motivation for including this additional cost is that limiting relationships with other countries will generate some welfare loss; to illustrate the impact of this modelling choice, we also analyze the limiting case where punishment can be inflicted at no cost.

Under our assumptions about the clauses included in the agreement, we first solve the dynamic emissions game for a given membership size. We subsequently use a dynamic version of the internal and external stability concept of d'Aspremont et al. (1983) to define a stable agreement. Using this stability concept, we then solve the (dynamic) membership game, so that the stable size of the agreement is obtained, as a function of the level of the pollution stock. Finally, we propose an evolutionary process that may describe how countries reach a stable situation, where the level of the pollution stock and the size of the IEA membership do not change anymore.

Here it is important to stress the difference between our assumptions on how countries reach a stable IEA and other dynamic game models. In Rubio and Ulph (2007), for instance, it is assumed that, at each decision epoch, countries solve both an emissions game and a membership game. Thus, at each epoch, given the level of the pollution stock, the size of the stable coalition is computed, and a group of signatory countries is randomly selected. This process evolves towards a steady-state of the pollution stock to which corresponds a steady-state for the size of a stable IEA.

In this paper, we do not assume that renegotiations between countries to decide whether or not to be a member of the agreement occur at every period. Negotiations only take place at an "initial" moment, but the membership of the agreement is not closed, and countries can decide at any decision epoch to join or to quit the coalition, following an evolutionary process based on individual welfare considerations. This process evolves towards a steady-state of the pollution stock to which corresponds a steady-state for the size of a stable IEA.

The main strength of an evolutionary process is that it is able to depict the stickiness and delays that may characterize this convergence. The motivation for a gradual convergence to stability finds its justification in practice: often, an IEA is promoted and signed by a first group of founding countries (which are, for some reason, more sensitive to the environmental problem) and then, over time, other countries come to sign the agreement.[4] An example is the Montreal Protocol, which, after being ratified by a first group of countries, was joined by new members every year.

[3] It however has been considered in the fishery context (see, e.g., Sethi and Somanathan 1996 and Bischi et al. 2004).

[4] Of course, it may also happen that a member country decides to abandon the agreement.

In particular, we adopt replicator dynamics in discrete time which provide evolutionary pressures in favor of the group obtaining the highest payoff. Following the spirit of evolutionary games, the group that performs better is joined by a fraction of new agents. The adjustment speed at which countries switch to the superior strategy is related to the difference in welfare, and it reflects the intangible or political cost of changing behaviors. It is worthwhile noticing that, in this process, the evolution of the players' welfare over time depends not only on the dynamics of emissions and pollution, but also on how the composition of the different groups evolves.

Our model is able to capture different long-run situations: where no stable agreement is reached, where all countries join the agreement, and (more realistically) where some countries take part in a stable IEA, and others do not.

The paper is organized as follows. Section 2 presents the model and the general dynamics that governs the evolution of the pollution stock. In Section 3, we propose and solve a dynamic game in which countries optimize their welfare over an infinite horizon by taking into account the evolution of the pollution stock. Section 4 introduces the stability concept, solves the corresponding membership game, and proposes a replicator dynamics for the number of signatory countries. Section 5 presents numerical illustrations and various analyses. Finally, Section 6 concludes the paper. Computational details are provided in the Appendix.

2 The Model

Let us consider N identical countries. A fraction s of them, identified as "signatory countries," decides to join an international environmental agreement, according to which their production activity is decided by maximizing the aggregate welfare of the coalition. We denote by S the set of signatory countries. The remaining fraction $(1-s)$, identified as "non-signatory countries" or "defectors," acts individually, that is, each of them decides its production activity by maximizing its individual welfare, and we denote by D the set of defectors. Note that this assumption on the behavior of signatories and non-signatories is common to both the cooperative and noncooperative approaches when they deal with IEAs.

Each country has a production activity that generates benefits but also pollution. Let us indicate with e_{jt} the emissions generated by the production of country j in time period t. We suppose that the net revenue (i.e., gross revenue minus costs) derived from country j's production activity in a given period is an increasing concave function of its emissions, and given by the quadratic function

$$R(e_{jt}) = e_{jt}\left(b - \frac{1}{2}e_{jt}\right).$$

Countries suffer an environmental damage arising from (global) pollution, which is assumed linear and given by

$$D(P_t) = dP_t,$$

where $d > 0$ is the constant marginal damage and P_t is the stock of pollution at time t.

As part of the agreement, we assume that each signatory country has to punish a non-signatory for its irresponsible behavior (Hoel and Schneider 1997). Moreover, we suppose that this punishment is directly proportional to the level of pollution, reflecting a level of environmental concern by the signatories, that is increasing with pollution stock, so that the non-environmental cost incurred by a defector punished by Ns signatories is given by

$$\phi(s, P_t) = Ns\alpha P_t.$$

This sanction can be interpreted as a limitation of trade with a non-signatory country or a carbon tax imposed on its exports.

As punishing a country has some negative effect on signatory countries, we suppose that punishing itself has a cost, which is proportional to the punishment αP_t imposed to the $N(1-s)$ non-signatory countries,[5] so that each signatory incurs a non-environmental cost given by

$$\omega(s, P_t) = N(1-s)\tau\alpha P_t,$$

where $\tau \geq 0$. As a consequence, the welfare of a signatory country $j \in S$ in time period t is given by

$$W_t^S(e_{jt}, P_t, s) = e_{jt}\left(b - \frac{e_{jt}}{2}\right) - dP_t - \tau\alpha N(1-s)P_t,$$

whereas the welfare of a non-signatory $j \in D$ is given by

$$W_j^D(e_{jt}, P_t, s) = e_{jt}\left(b - \frac{e_{jt}}{2}\right) - dP_t - \alpha NsP_t.$$

In the sequel, we will use the following convenient abbreviated notation

$$c_S \equiv d + \tau\alpha N(1-s) \quad (1)$$

$$c_D \equiv d + \alpha Ns \quad (2)$$

to represent the marginal (environmental and non-environmental) impacts of pollution on the welfare of signatory and non-signatory countries, respectively (which are however not constants, but linear functions of s). Notice that the difference in marginal costs $c_S - c_D$ is a linear function of s which is positive for $s < \frac{\tau}{1+\tau}$ and negative for $s > \frac{\tau}{1+\tau}$.

[5] The marginal cost of punishing is assumed to be independent from the number of signatory countries. This is the case, for example, of a trade sanction: the cost to a member's welfare depends only on the number of countries to which the sanctions are applied.

Finally, the evolution over time of the pollution stock is assumed to be governed by the discrete time equation

$$P_t = P_{t-1}(1-\delta) + \sum_{i \in S} e_{it} + \sum_{k \in D} e_{kt} \tag{3}$$

where $\delta \in (0,1)$ is the natural decay, $\sum_{i \in S} e_{it}$ is the total emissions of signatory countries and $\sum_{k \in D} e_{kt}$ is the total emissions of non-signatories at time period t.

This model is a deliberate simplification of reality, which allows us to derive closed-form solutions and theoretical properties. It however accounts for two important features of the environmental agreement stability problem: first, the dynamics of the pollution stock and of the related damage cost, and, second, the negative externalities arising from global pollution, that is, from the emissions of all players.

The assumption of identical players is not a crucial one. Indeed, the model can easily be extended to the case of asymmetric countries, and the equilibrium strategies are readily obtained - but one has then to provide specific values for the parameters (see for instance Bahn et al. 2008 for an illustration of this model, calibrated to nine different regions using the MERGE policy assessment model).

The assumption of a quadratic revenue function is rather popular in the environmental economics literature (see, e.g., Rubio and Casino 2005; Diamantoudi and Sartzetakis 2006; Rubio and Casino 2002; Dockner and Long 1993), and captures the diminishing returns of production activity. The most debatable assumption is clearly the linear environmental damage. This simplification is not uncommon (see for instance Hoel and Schneider 1997; Finus 2004) and is supported by some empirical estimations (see Labriet and Loulou 2003). It can be motivated by the assumption that players approximate the damage function using a local marginal information. However, it ignores potential catastrophic consequences that could happen with significant increases in the pollution stock. A convex damage function is a more realistic choice, at the expense of tractability; using a quadratic damage function in our model yields qualitatively similar outcomes, no longer in closed-form. In the Conclusion, we discuss further the impact, on both the qualitative results and the tractability of the problem, of adopting a quadratic damage function rather than a linear one.

3 A Dynamic Game of Emissions

In this section, we solve the dynamic emissions game. We assume that, for a given fixed number of signatories, countries optimize their welfare by taking into account the evolution of the pollution stock. The total discounted welfare of players is maximized over an infinite horizon, where $\beta \in (0,1)$ is the one-period discount factor assumed common to all players.

The welfare maximization problem for a signatory country $j \in S$ is thus given by

$$\begin{aligned} \max_{(e_j), j\in S} W^S &= \sum_{j\in S}\sum_{t=0}^{\infty} \beta^t \left(e_{jt}\left(b - \frac{e_{jt}}{2}\right) - P_t c_S \right) \\ &\text{s.t.} \\ P_t &= P_{t-1}(1-\delta) + \sum_{i\in S} e_{it} + \sum_{k\in D} e_{kt}, P_0 \text{ given}, \end{aligned} \tag{4}$$

where e_{jt} is the emissions of country j during period t and e_j denotes the sequence of emissions $\{e_{jt}\}_{t=0,\dots\infty}$. In the same way, the welfare maximization problem for a defector country $j \in D$ is

$$\max_{e_j} W^D = \sum_{t=0}^{\infty} \beta^t \left(e_{jt}\left(b - \frac{e_{jt}}{2}\right) - P_t c_D \right),$$

subject to (4).

In order to characterize the players' optimal reaction strategies and the dynamic equilibrium, we use a dynamic programming formulation where the state variable is P, that is, the pollution stock level in the preceding time period. We solve for a Nash equilibrium in stationary feedback strategies between the group of signatories, acting as a single player, and the non-signatories, acting as $N(1-s)$ individual players, where $Ns \in [1, N-1]$. We call κ the constant representing the combined effect of the discount factor and the natural pollution decay,[6] that is,

$$\kappa \equiv \frac{1}{1-\beta(1-\delta)} > 1.$$

We first compute the optimal reaction of the set of signatory countries to a given stationary strategy vector for the defectors, denoting by $E^D(P)$ the resulting total emissions of non-signatory countries as a function of P. For each signatory country, the value function $V^S(P; E^D)$ represents the optimal total welfare of a signatory country, given $E^D(P)$, and it satisfies

$$\begin{aligned} V^S(P; E^D) = \max_e \Big\{ & e\left(b - \frac{e}{2}\right) - \left(P(1-\delta) + Nse + E^D(P)\right) c_S \\ & + \beta V^S\left(P(1-\delta) + Nse + E^D(P); E^D\right) \Big\}. \end{aligned} \tag{5}$$

Proposition 1 *The value function of a signatory country is linear in P. The optimal reaction of signatory countries is independent of the level of pollution and of the defectors' strategy, and is given by*

$$e^S = b - Ns\kappa c_S, \tag{6}$$

assuming $b > Ns\kappa c_S$.

[6] A value of $\kappa = 1$ corresponds to myopic limiting cases, with either $\beta = 0$ (no value for the future) or $\delta = 1$ (no stock accumulation).

Proof. Assume that $V^S(P; E^D) = k^S - m^S P$. Then, first-order sufficient conditions yield

$$e^S = b - Ns\left(m^S\beta + c_S\right),$$

which achieves the maximum in (5) and which does not depend on P or $E^D(P)$. Replacing e^S in (5) we obtain an expression that is linear in P and that verifies our assumption. It is now straightforward to get by identification (see Appendix 7.1)

$$\begin{aligned} m^S &= c_S\kappa(1-\delta) \\ k^S &= Ns\kappa c_S\frac{Ns\kappa c_S - 2b}{2(1-\beta)} + \frac{b^2}{2(1-\beta)} - \frac{\kappa E^D(P)\, c_S}{(1-\beta)}. \end{aligned}$$

The optimal emissions of a signatory country are therefore

$$e^S = b - Ns\kappa c_S.$$

■

Notice that the emissions of signatory countries endogenize the damage of the entire coalition and are convex in s. It is then straightforward to compute the total emissions of signatory countries as $E^S = Ns(b - Ns\kappa c_S)$.

In the same way, we now express the optimal reaction of a defector to a given stationary-strategy vector of the other countries, denoting by $E^d(P)$ the total emissions of all other non-signatory countries as a function of P, and by E^S the total emissions of the signatory countries. The value function V^D of a defector country represents the optimal total welfare of a defector, given P, $E^d(P)$ and E^S, and it satisfies

$$\begin{aligned} V^D(P; E^S, E^d) &= \max_e \left\{ e\left(b - \frac{e}{2}\right) - \left(P(1-\delta) + E^S + e + E^d(P)\right) c_D \right. \\ &\quad \left. + \beta V^D\left(P(1-\delta) + E^S + e + E^d(P); E^S, E^d\right)\right\}. \end{aligned} \tag{7}$$

Proposition 2 *The value function for a defector country is linear in P. The optimal reaction of non-member countries is independent of the level of pollution and of the other players' strategy, and is given by*

$$e^D = b - \kappa c_D, \tag{8}$$

assuming $b > \kappa c_D$.

Proof. Assume that $V^D(P; E^S, E^d) = k^D - m^D P$. Then, first-order sufficient conditions yield

$$e^D = b - c_D - m^D\beta.$$

Substituting e^D in (7) we obtain a linear function of P and (see Appendix 7.2)

$$\begin{aligned} m^D &= c_D\kappa\left(1-\delta\right) \\ k^D &= \frac{b^2-\kappa c_D\left(2b-\kappa c_D+2\left(E^S+E^d\left(P\right)\right)\right)}{2\left(1-\beta\right)}, \end{aligned}$$

so that the optimal emissions for a defector country are

$$e^D = b-\kappa c_D.$$

■

Notice that the optimal emissions[7] of a defector are linear decreasing in s, independent of P and of the strategies of the other players. It is then straightforward to compute the total emissions of the other defector countries as $E^d = (N-Ns-1)(b-\kappa c_D)$.

Combining these results, the equilibrium strategy pair for both kinds of players is given by $\left(e^S, e^D\right)$ and the total emissions at equilibrium are

$$E(s) \equiv N\left(b-\kappa\left(c_D\left(1-s\right)+Ns^2c_S\right)\right).$$

The consequence of assuming a constant marginal environmental damage is that the optimal emissions of countries are independent of each other (orthogonal free-riding), but they are still linked because they are both functions of s.

It is straightforward to verify that the full defection equilibrium, where all players maximize their individual welfare, that is,

$$\max_{e_j} W^D = \sum_{t=0}^{\infty}\beta^t\left(e_{jt}\left(b-\frac{e_{jt}}{2}\right)-P_td\right),$$

subject to (4), can be obtained by setting s equal to 0 in (7). Similarly, the cooperative solution, where all players agree to maximize their aggregate welfare, that is,

$$\max_{(e_j)} W^S = \sum_{j=1}^{N}\sum_{t=0}^{\infty}\beta^t\left(e_{jt}\left(b-\frac{e_{jt}}{2}\right)-P_td\right)$$

subject to (4), can be obtained by setting s equal to 1 in (5). We extend the domain of the value functions in the obvious way: when $s=0$, V^D represents the total discounted welfare and e^D the equilibrium strategy of players when there is no coalition, and when $s=1$,

[7] The emissions of a defector are higher than those of a signatory for $Ns \in [1, N-1]$ iif $d > \alpha$ and $\tau > \frac{1}{N-1}$ (see Appendix 7.3). Condition $d > \alpha$ is satisfied in our numerical illustrations. This means that the punishment inflicted on a non-member country by an individual signatory is small with respect to the damage it suffers from pollution. If countries were able to inflict very large punishments at very small costs, then they could make the non-signatories pollute even less than the cooperative optimum.

V^S represents the total discounted welfare and e^S the optimal strategy of players under full cooperation.

Replacing (6) and (8) in (3), the dynamics of the pollution stock becomes

$$P_t = P_{t-1}(1-\delta) + N\left(b - \kappa\left(c_D(1-s) + Ns^2 c_S\right)\right). \tag{9}$$

For a given s, the steady state of the pollution stock is

$$P^*(s) = \frac{E(s)}{\delta}, \tag{10}$$

and the individual total discounted welfares of a signatory and non-signatory country when the pollution stock is P are, respectively (see Appendix 7.4),

$$V^S(s,P) = \frac{b^2 - N^2\kappa^2 c_S^2 s^2}{2(1-\beta)} - \kappa c_S\left(P(1-\delta) + \frac{E(s)}{1-\beta}\right) \tag{11}$$

$$V^D(s,P) = \frac{b^2 - \kappa^2 c_D^2}{2(1-\beta)} - \kappa c_D\left(P(1-\delta) + \frac{E(s)}{1-\beta}\right). \tag{12}$$

4 Evolution and Stability of IEAs

We now consider a dynamic version of the stability concept introduced by d'Aspremont et al. (1983) and widely used in static games. We first solve the membership game under this stability concept, that is, we state the necessary conditions characterizing a stable coalition, for a given level of the pollution stock. We then propose a corresponding discrete-time replicator dynamics for the proportion of signatories, that will allow players to reach a stable IEA in the long run.

4.1 Unilateral deviation and foresight

In a static context, the internal and external stability concept assumes that each player is able to compare his welfare with what he would achieve if he unilaterally switched group. In a dynamic context, where players' criterion is total discounted welfare, the conditions for internal and external stability become, respectively,

$$V^S(s,P) \geq V^D\left(\frac{Ns-1}{N}, P\right) \tag{13}$$

$$V^S\left(\frac{Ns+1}{N}, P\right) < V^D(s,P). \tag{14}$$

Inequality (13) means that, at P, the total discounted welfare over an infinite horizon of a country in a coalition of Ns signatories is larger than what it would attain as a defector against a coalition of $Ns-1$ countries – so that there is no incentive for a signatory to defect. Inequality (14) means that, at P, the total discounted welfare over an infinite horizon of a

defector country against a coalition of Ns signatories is larger than what it would attain as a signatory in a coalition of $Ns+1$ countries – so that there is no incentive for a defector to join the coalition.

Define the stability function

$$\Psi(s,P) = V^S(s,P) - V^D(\frac{Ns-1}{N},P).$$

It is easy to check (see Appendix 7.5) that the extension of $\Psi(s,\cdot)$ to $\left[\frac{1}{N},1\right]$ is continuous in s, and that a necessary condition for IEA stability with $\lfloor Ns \rfloor$ signatories when the stock of pollution is $\bar{P}$ is obtained when $\Psi(s,\bar{P})=0$ and it is given by

$$\begin{aligned}\bar{P}\left(1-\beta\right)\left(1-\delta\right) &= -Nb+\kappa\frac{\left(Ns-1\right)^2\left(c_S+\alpha\tau\right)\left(c_D-\alpha\right)}{\left(c_D-\alpha-c_S\right)} \\ &\quad +\kappa\frac{\left(2N\left(1-s\right)+1\right)\left(c_D-\alpha\right)^2-Nc_S\left(2c_D\left(1-s\right)+Ns^2c_S\right)}{2\left(c_D-\alpha-c_S\right)}. \quad (15)\end{aligned}$$

Note that (15) characterizes the solution of the membership game. For any level of pollution, it provides a condition that must be satisfied by the size of a stable coalition.

4.2 Dynamics

As the pollution stock level changes over time, the number of signatories required for a stable IEA also changes. In order to describe this dynamics, one could follow the idea proposed in Rubio and Ulph (2007) for instance, where, in each period, countries solve a membership game. This amounts to finding the number of signatories satisfying the stability condition (15) at P_t; in practice, this would require at each time period either an exogenous intervention or some time-consuming negotiation process.

Instead, we introduce a mechanism that describes how countries may gradually reach a stable IEA.[8] Assume that the proportion of signatory countries evolves over time following the discrete time replicator dynamics[9]

$$s_{t+1} = s_t\frac{V^S(s_t,P_t)}{s_tV^S(s_t,P_t)+(1-s_t)V^D(s_t-\frac{1}{N},P_t)} \text{ if } s_t\in[1/N,1], \quad (16)$$

where V^S and V^D are given by (11)-(12) and assumed positive[10].

[8] This means that, while the level of the pollution stock is not stabilized, the size of the coalition is not the one resulting from a membership game.

[9] Replicator dynamics in continuous time is commonly used in the context of common pool resource games and, in particular, in the fisheries, to describe the evolution of a population of agents where two behaviors can be adopted (see Sethi and Somanathan 1996; Noailly et al. 2007; Osés-Eraso and Viladrish-Grau 2007 for instance). In Xepapadeas and Passa (2004), the replicator dynamics is used to model the participation and compliance of firms to voluntary environmental agreements.

[10] It is straightforward to adapt the adjustment mechanism if the welfare of one or both groups of players is negative, or if the proportion of signatories vanishes (see Appendix 7.6).

The denominator of (16) represents a weighted average of the (total discounted) welfare of a signatory country, and of a country which unilaterally defects from the agreement. The weights are given by the current proportions of signatories and defectors.

Whenever the welfare obtained from one type of behavior is higher than the other, a fraction of countries will switch behavior to join the group that is performing better. Equation (16) captures the notion that a strategy yielding profits above *(below)* the average increases *(decreases)* the relative share of the population using that strategy[11] and that the "speed" of change depends on the relative inequalities of welfare. This update mechanism ensures that any shift in population shares is a gradual process, and shows that there might be some delays or inertia involved in countries readjusting their behavior, for instance, due to an intangible or political cost of switching.

In this case, (9) and (16) give rise to a two-dimensional dynamic system that describes the evolution, over time, of the stock of pollution and of the proportion of signatories. It is easy to check that a steady state (s^*, P^*) of this dynamic system satisfies the stability condition for an IEA at $\bar{P} = P^*(s^*)$ (see Appendix 7.6). Players reach a stable IEA by comparing total discounted welfares when staying in or unilaterally leaving a coalition, without taking into account the dynamics of s.

5 Steady-State and Sensitivity Analysis

We now study the dynamics of the pollution and of countries' shares when the players behavior is given by (16). In particular, we are interested in finding out whether or not full cooperation, the coexistence of cooperators and defectors, and no cooperation at all are all possible outcomes of this model, and under what conditions they would eventually occur. In our numerical simulations, we use parameter values that give rise to positive individual emissions.

When they exist, equilibrium steady-state values of the stock of pollution and proportion of signatory countries are indexed by υ and we denote $\xi_\upsilon = (s_\upsilon, P_\upsilon)$ an equilibrium steady state of the dynamic system (9)–(16).

We first notice that the boundary equilibria ξ_n, corresponding to $s_n = 0$, and ξ_c, corresponding to $s_c = 1$, may be reached by the dynamic system. These are given by

$$\begin{aligned} \xi_n &= \left(0, \frac{N\left(b - \kappa d\right)}{\delta}\right) \\ \xi_c &= \left(1, \frac{N\left(b - \kappa dN\right)}{\delta}\right), \end{aligned}$$

[11] This idea is not uncommon in economics. For example, when new strategies or technologies are introduced on the market, firms will tend to imitate the most successful ones, or the ones that yield a 'satisfactory' level of profits. Here, we assume that governments follow the same kind of behavior.

where $P_n > P_c$. The dynamic system may also reach inner steady states corresponding to IEAs with partial cooperation.

5.1 Partial cooperation

According to the parameter values, 0, 1 or 2 coexisting inner steady states may appear, corresponding to situations where the necessary condition for stability is satisfied at the steady-state pollution. These are defined by the intersections of $P^*(s)$ with $\bar{P}(s)$. When two inner steady states exist, the one with the lower percentage of signatory countries, denoted by ξ_l, is a saddle point[12] and the one with the higher percentage of signatory countries, denoted by ξ_h, is a stable node, corresponding to a stable IEA with partial cooperation. Figures 1–3 illustrate in the (s, P) plane three representative examples, and how these situations may appear when the values of parameters α and τ are changed.

Figure 1 illustrates the impact of an increase in the punishment α. In Figure 1a (with $\alpha = 0.00029$), functions P^* and $\bar{P}$ do not intersect: there is no value of s such that the stability criterion is satisfied at the steady-state pollution stock. From any initial conditions (s_0, P_0), the dynamic system converges to ξ_n, that is, no stable agreement is possible. In Figure 1b (with $\alpha = 0.000305$), if the initial group of signatory countries is large enough (that is, if the initial conditions are in the blue area), then the steady-state ξ_h is reached, corresponding to a stable IEA with partial cooperation. An interesting observation is that the minimum size of the initial number of signatories that leads to the inner steady state is decreasing with the initial pollution stock (see Figure 2 for a zoom-in). This means that if the level of accumulated pollution is high, only a small number of founding countries is necessary to reach a stable IEA.

In general, an increase of α decreases both $P^*(s)$ and $\bar{P}(s)$; it determines a higher level of cooperation, a lower level of pollution at the steady state, and it enlarges the basin of initial states generating trajectories converging to ξ_h. In Figure 1c (with $\alpha = 0.000337$), starting from any initial condition in the green area, the system converges to ξ_c. Further increases in punishment do not have any effect on the long-run values of the dynamic variables, because full cooperation has been reached, but they make full cooperation more robust (i.e., supported by a greater number of initial states).

Figure 3 illustrates the impact of decreasing τ, starting from the case depicted in Figure 1a (with $\tau = 0.65574$). A decrease in τ decreases both P^* and $\bar{P}$. In Figure 3a (with $\tau = 0.5862$), an inner steady state with partial cooperation appears. Further decreases in τ have positive effects on the long-run values of the dynamic variables (higher cooperation and lower pollution stock), and especially, on the basin of the stable steady state (Figure 3b). This means that, if the mandatory punishment set out by the agreement is not too expensive, the initial group of founding countries does not need to be very large to lead to a stable agreement. Reducing

[12] A saddle point corresponds to a root where the stability function is increasing. Entry and exit conditions are not satisfied at a saddle point.

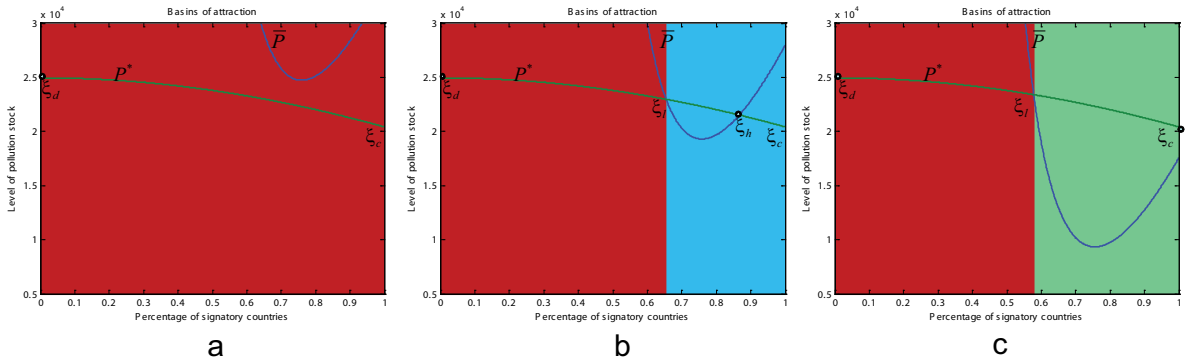


Figure 1: Impact of increasing punishment. The parameter values are: $N = 100$, $b = 200$, $d = 0.3$, $\tau = 0.65574$, $\delta = 0.8$, $\beta = 0.9$. In (a), $\alpha = 0.00029$; in (b), $\alpha = 0.000305$; in (c), $\alpha = 0.000337$. The red area represents the set of initial conditions of s and P generating trajectories converging to the fully non-cooperative outcome. The blue area represents the set of initial conditions of s and P generating trajectories converging to the partial cooperation outcome. The green area represents the set of initial conditions of s and P generating trajectories converging to the full cooperative outcome.

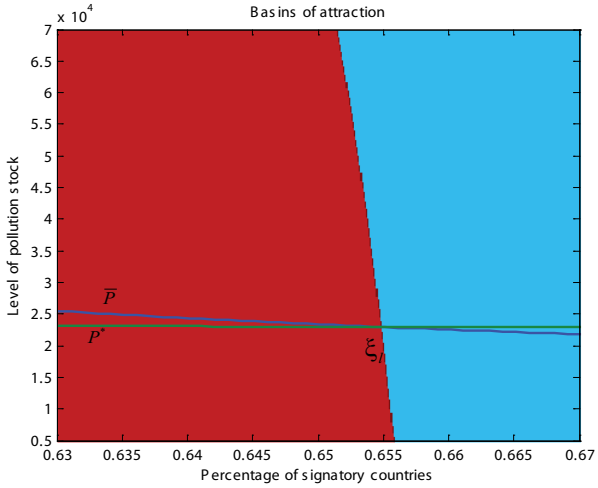
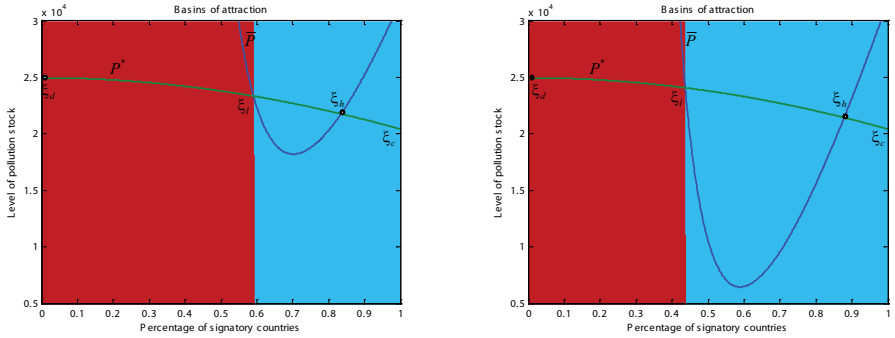


Figure 2: Zoom in on the basin of attraction.

the cost for punishing may or may not generate full cooperation, depending on the value of the other parameters (this is discussed further in Section 5.2).

A sensitivity analysis with respect to other parameters shows that increasing the profitability of emissions (parameter b) has a positive impact on the number of signatories at the steady state, as well as on the set of initial states generating trajectories converging to partial or full cooperation; however, this is at the expense of a higher steady-state pollution stock (this is due to the fact that, even if the number of cooperators increases, they also have a higher incentive to emit more). More interestingly, an increase in the marginal environmental cost (parameter d) negatively impacts on the size of the steady-state coalition as well as on



a b

Figure 3: Impact of decreasing cost for punishment. The parameter values are: $N = 100$, $b = 200$, $d = 0.3$, $\alpha = 0.00029$, $\delta = 0.8$, $\beta = 0.9$. In (a), $\tau = 0.5862$; in (b), $\tau = 0.4483$. The red area represents the set of initial conditions of s and P generating trajectories converging to the fully non-cooperative outcome. The blue area represents the set of initial conditions of s and P generating trajectories converging to the partial cooperation outcome.

the size of the basin of attraction leading to it. This result is consistent with what has been observed in static games: when the potential gain from cooperation is high, the membership in an IEA is likely to be small. This can be explained by the incentive to free ride, implying that a stable IEA is the smallest where emitting less is welfare-enhancing.

Figure 4 represents various trajectories for the pollution stock and the number of signatory countries over time for the case depicted in Figure 1b, where an inner steady state exists. Depending on the initial conditions, trajectories converge either to a situation with no cooperation or to a stable IEA. The way this stable solution is reached also depends on the initial condition. Possible evolutions in the space (s, P) are illustrated in Figure 5. Notice how the evolutionary pressure increases s above $\bar{P}$ (where signatories fare better) and decreases s otherwise.

5.2 No cost for punishing

We now consider the special case when enforcing a sanction does not entail any cost (as in Hoel and Schneider 1997). In this case, the marginal cost of pollution for a non-signatory country is always higher than that of a signatory country, and a non-signatory emits more than a signatory for $Ns\,(d - \alpha) > d$.

In addition, one can show by studying the stability function that the entry and exit conditions can be satisfied at at most one point (see Appendix 7.7); if punishing has no adverse consequences on the welfare of coalition members, then there is at most one inner equilibrium with coexistence of defectors and signatories. Moreover, the system never admits as a solution the case of complete defection, but the dynamic system converges either to an inner steady state with partial cooperation, or to a situation with full cooperation.

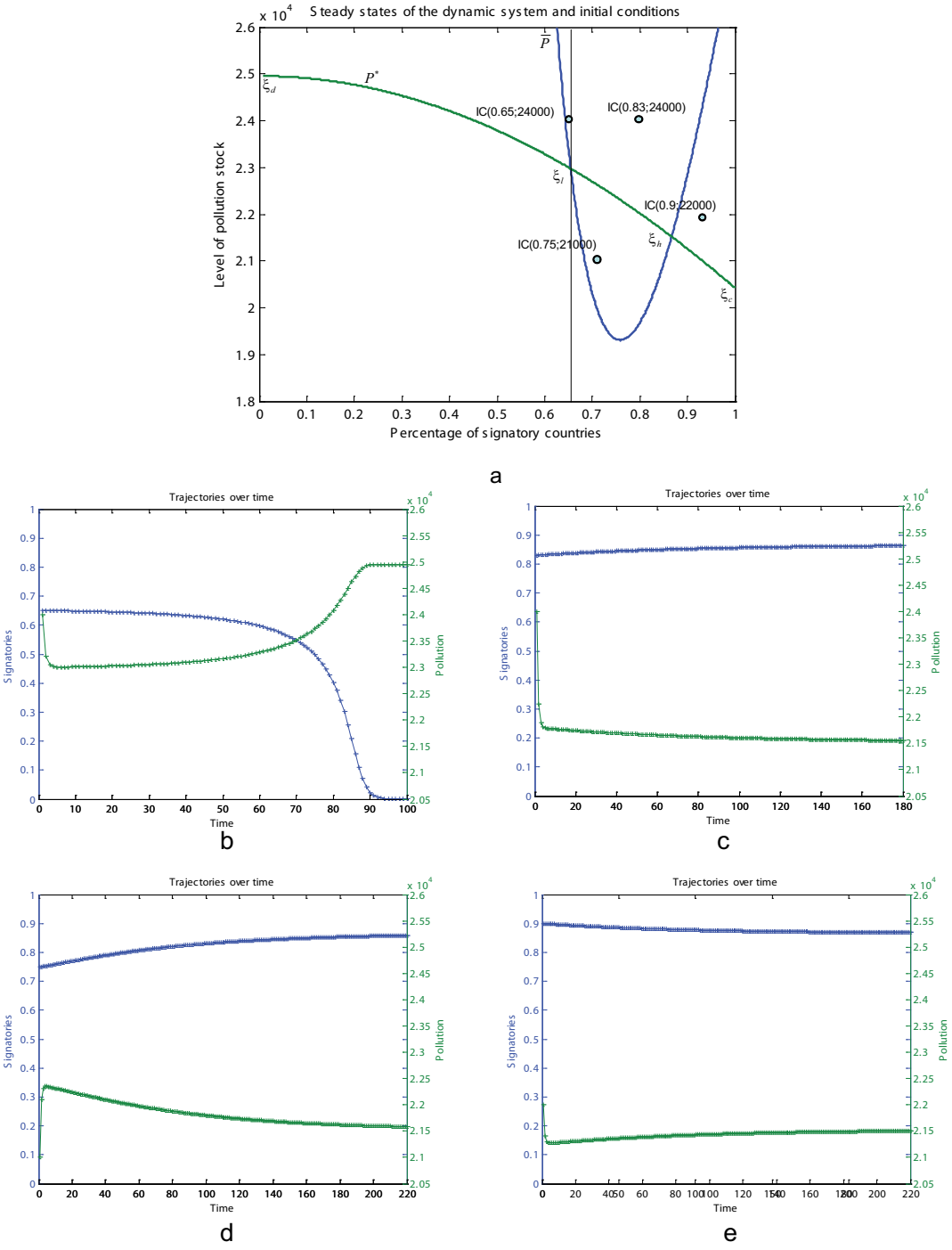


Figure 4: Trajectories corresponding to various initial conditions. Parameters are as in Figure 1b. In panel b, the number of signatories converges to 0. In panels c, d and e, trajectories converge to the same inner steady state, but following different paths depending on the initial conditions.

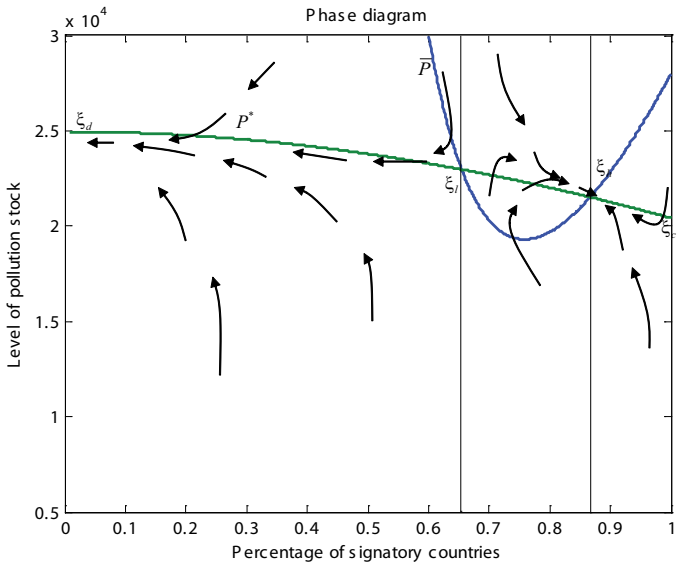


Figure 5: Evolution of s and P.

5.3 No punishment

An interesting question is how the dynamic model performs when there is no punishment clause in the agreement ($\alpha = \tau = 0$). In this case, the exit condition

$$V^S(s,P) \geq V^D(\frac{Ns-1}{N},P)$$

reduces to $(Ns-3)(Ns-1) \leq 0$, while the entry condition

$$V^S(\frac{Ns+1}{N},P) < V^D(s,P)$$

reduces to $Ns > 2$. As a consequence, for any level of pollution, the only stable IEA involves 3 members, which corresponds to the result obtained for static games (see Barrett 1994).

5.4 Endogenous punishment

In the above analysis, we assumed that the punishment level agreed on by the signatories was exogenous. Recognizing that the best outcome is full cooperation, it is however possible for the initial group of signatories to compute a value for α such that full cooperation is achieved in the long run. The smallest level of α for which full cooperation is achieved at the steady state of pollution satisfies (15) at $\bar{P} = N\left(\frac{b-Nd\kappa}{\delta}\right)$ and $s = 1$, or equivalently solves (see Appendix 7.7)

$$\begin{aligned}&\alpha^2\kappa^2\delta(N-1)(2\tau(N-1)+1)\\ &+2\alpha\left(-Nb+d\kappa\left(\kappa\delta(\tau-2)(N-1)+N^2\right)\right)\end{aligned}$$

$$+d^2\kappa^2\delta\,(N-3) \quad = \quad 0.$$

For instance, with the parameter values used in Figure 1, one obtains $\alpha^* = 3.1693 \times 10^{-4}$.

6 Conclusions

In this paper, we dealt with the problem of stability of international environmental agreements concerning pollution emissions. Stock externalities were included, as well as the possibility for countries to abide by the agreement or to defect at any time. We considered an agreement that includes a provision for signatory countries to punish non-signatory countries, even if sanctioning the defectors entails a cost. We developed a model in which countries optimize their welfare over an infinite horizon, taking into account the evolution of the stock of pollution. In defining a stable coalition, we applied the well-known internal and external stability conditions (d'Aspremont et al. 1983). Finally, we proposed an evolutionary mechanism that might be used to reach a stable IEA over time. This idea is validated by actual practice where the design and "growth" of an IEA are experienced. A motivating example is the Montreal Protocol, to which many countries have adhered, in the 20 years since it was first ratified. This shows that IEAs with high participation do occur, and that membership in the agreement may change over time, without formal renegotiation.

The main results of our numerical simulations can be summarized as follows. The outcome in which no country joins the IEA is always a solution, but, provided that sanctions are strong enough and/or that the cost for punishing is not too high, this outcome coexists with either a partial-cooperation solution or with full cooperation. When two outcomes coexist, initial conditions are decisive: if the initial coalition is not large enough for a given initial level of pollution stock, then the equilibrium solution is full defection. The basin boundary can be interpreted as a minimum participation clause, where the minimum number of members is affected by the level of punishment included in the agreement and/or by its cost. However, the greater the initial level of pollution, the fewer initial signatory countries are needed to converge to a stable agreement, indicating that it is easier to endogenously bring other countries into the IEA when the global environmental damage is high. Finally, as in the static models, the number of signatories in a stable IEA is negatively related to the environmental cost or, equivalently, to the benefits of cooperation. This is to be expected from self-enforcing agreements where stability is attained when the welfares of signatories and defectors are close: when the benefits from cooperation are higher, the welfare of signatories rises more rapidly with the size of the coalition, and it takes less signatories to attain the welfare of defectors.

To complete the analysis, we considered the special case when enforcing a sanction doesn't entail any cost. In this case, full defection can never be observed; the only possible outcome, independent of the initial conditions but depending on the value of the parameters, is either

partial or full cooperation. We also considered the special case where punishment is not part of the agreement, and we found that stable coalitions with large numbers of players cannot be implemented, irrespective of the pollution stock. Finally, we observed that a punishment level ensuring that full cooperation will be reached in the long run can be easily obtained.

Our assumption of linear environmental damage makes the problem tractable, since the strategies of the players become independent of the state and of the other players' actions. A linear damage may be interpreted as a local approximation of the damage function by players. When a quadratic damage function is used, we can still show that the equilibrium value function is quadratic, and that the equilibrium strategies of the players are linear (both in the pollution stock and in the emissions of the other players). However, the equilibrium solution can no longer be obtained in closed-form, because the coefficients in the equilibrium strategies have to be obtained numerically. Our experiments using a quadratic damage function yielded qualitatively similar results.

References

d'Aspremont C, Jacquemin A, Gabszewicz JJ, Weymark JA (1983) On the stability of collusive price leadership. The Canadian Journal of Economics / Revue canadienne d'économique 16(1):17–25

Bahn O, Breton M, Sbragia L, Zaccour G (2008) Stability of international environmental agreements: An illustration with asymmetrical countries. International Transactions in Operational Research (to appear - DOI: 10.1111/j.1475-3995.2008.00678)

Barrett S (1994) Self-enforcing international environmental agreements. Oxford Economic Papers 46:878–94

Barrett S (1997) The strategy of trade sanctions in international environmental agreements. Resource and Energy Economics 19:345–361

Barrett S (2003) Increasing participation and compliance in international climate change agreements. International Environmental Agreements: Politics, Law and Economics 3:349–376

Bischi GI, Lamantia F, Sbragia L (2004) Competition and cooperation in natural resources exploitation: An evolutionary game approach. In: Carraro C, Fragnelli V (eds) Game Practice and the Environment, E. Elgar Publishing, Cheltenham, p 187–221

Botteon M, Carraro C (1998) Strategies for environmental negotiations: Issue linkage with heterogeneous countries. In: Hanley N, Folmer H (eds) Game Theory and the Environment, E. Elgar Publishing, Cheltenham, p 181–203

Cabon-Dhersin ML, Ramani S (2006) Can social externalities solve the small coalitions puzzle in international environmental agreements? Economics Bulletin 17:1–8

Carraro C, Marchiori C, Orefice S (2003) Endogenous minimum participation in international environmental treaties. Working paper No. 113.2003 FEEM

Carraro C, Siniscalco D (1993) Strategies for the international protection of the environment. Journal of Public Economics 52:309–328

Carraro C, Siniscalco D (1997) R&D cooperation and the stability of international environmental agreements. In: Carraro C (ed.), International Environmental Negotiations, Strategic Policy Issues, E. Elgar, Cheltenham, p 71–96

Carraro C, Siniscalco D (1998) International environment agreements: Incentives and political economy. European Economic Review 42:561–572

Diamantoudi E, Sartzetakis ES (2006) Stable international environmental agreements: An analytical approach. Journal of Public Economic Theory 8(2):247–263

Dockner E, Jorgensen S, Long NV, Sorger G (2000) Differential Games in Economics and Management Science. Cambridge University Press

Dockner E, Long NV (1993) International pollution control: Cooperative versus noncooperative strategies. Journal of Environmental Economics and Management 24:13–29

Dockner E, Nishimura K (1999) Transboundary pollution in a dynamic game model. The Japanese Economic Review 50(4):443–456

Endres A (2004) Game theory and global environmental policy. Poiesis & Praxis 3:123–139

Finus M (2000) Game theory and international environmental co-operation: A survey with an application to the Kyoto Protocol. Nota di lavoro 86.2000, Fondazione Eni Enrico Mattei

Finus M (2004) Modesty pays: Sometimes. Nota di lavoro 68.2004, Fondazione Eni Enrico Mattei

Germain M, Toint P, Tulkens H, de Zeeuw A (2003) Transfers to sustain dynamic core-theoretic cooperation in international stock pollutant control. Journal of Economic Dynamics and Control 28:79–99

Hardin G (1968) The tragedy of the commons. Science 162(3859):1243–1248

Hoel M, Schneider K (1997) Incentives to participate in an international environmental agreement. Environmental and Resource Economics 9:153–170

Jeppensen T, Andersen P (1998) Commitment and fairness in environmental games. In: Hanley N, Folmer H (eds) Game Theory and the Environment, E. Elgar, Cheltenham, p 65–83

Katsoulacos Y (1997) R&D spillover, cooperation, subsidies and international agreements. In: Carraro C (ed) International Environmental Negotiations: Strategic Policy Issues, E. Elgar, Cheltenham, p 97–109

Labriet M, Loulou R (2003) Coupling climate damages and ghg abatement costs in a linear programming framework. Environmental Modeling & Assessment 8(3):261–274

Le Breton M, Soubeyran A (1997) The interaction between international environmental and trade policies. In: Carraro C (ed) International Environmental Negotiations – Strategic Policy Issues, E. Elgar Publishing, Cheltenham, p 126–149

Long NV (1992) Pollution Control: A differential game approach. Annals of Operational Research 37:283–296

Mohr E, Thomas J (1998) Pooling sovereign risks: The case of environmental treaties and international debt. Journal of Development Economics 55(1):153–169

Noailly J, Withagen CA, Van Den Bergh JCJM (2007) Spatial evolution of social norms in a common-pool resource game. Environmental & Resource Economics 36:113–141

Osés-Eraso N, Viladrich-Grau M (2007) On the sustainability of common property resources. Journal of Environmental Economics and Management 53:393–410

Rubio S, Casino B (2002) A Note on cooperative versus non-cooperative strategies in international pollution control. Resource and Energy Economics 24:251–261

Rubio S, Casino B (2005) Self-enforcing international environmental agreements with a stock pollutant. Spanish Economic Review 7:89–109

Rubio S, Ulph A (2006) Self-enforcing environmental agreements revisited. Oxford Economic Papers 58:233–263

Rubio S, Ulph A (2007) An Infinite-horizon model of dynamic membership of international environmental agreements. Journal of Environmental Economics and Management 54(3):296–310

Sethi R, Somanathan E (1996) The evolution of social norms in common property resource use. The American Economic Review 86(4):766–788

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

Wagner UJ (2001) The design of stable international environmental agreements: Economic theory and political economy. Journal of Economic Surveys 15(3):377–411

Xepapadeas A, Passa C (2004) Participation in and compliance with public voluntary environmental programs: An evolutionary approach. Nota di Lavoro 67.2004, Fondazione Eni Enrico Mattei

7 Appendix

7.1 Equilibrium emissions of signatory countries for fixed s

Assume that $V^S(P;E^D) = k^S - m^S P$. We then have

$$\begin{aligned} V^S(P;E^D) &= \max_e \Big\{ e\left(b-\frac{e}{2}\right) - \left(P(1-\delta) + Nse + E^D(P)\right)c_S \\ &\quad +\beta V^S\left(P(1-\delta) + Nse + E^D(P);E^D\right)\Big\} \\ &= \max_e \Big\{ e\left(b-\frac{e}{2}\right) - \left(P(1-\delta) + Nse + E^D(P)\right)c_S \\ &\quad +\beta\left(k^S - m^S\left(P(1-\delta) + Nse + E^D(P)\right)\right)\Big\}. \end{aligned}$$

Differentiating with respect to emissions yields:

$$\begin{aligned} \frac{d}{de}\big(e(b-e/2) - \left(P(1-\delta) + Nse + E^D(P)\right)c_S \\ +\beta\left(k^S - m^S\left(P(1-\delta) + Nse + E^D(P)\right)\right)\big) &= b - e - Ns\left(c_S + m^S\beta\right), \\ \frac{d}{de}\left(b - e - Ns\left(c_S + m^S\beta\right)\right) &= -1, \end{aligned}$$

so that the first order conditions are necessary and sufficient, provided the solution is interior. The FOC are satisfied at

$$e^S = b - Ns\left(m^S\beta + c_S\right).$$

Replacing e^S in (5), we obtain

$$\begin{aligned} V^S(P;E^D) &= \left(b - Ns\left(m^S\beta + c_S\right)\right)\left(b - \frac{\left(b - Ns\left(m^S\beta + c_S\right)\right)}{2}\right) \\ &\quad -\left(P(1-\delta) + Ns\left(b - Ns\left(m^S\beta + c_S\right)\right) + E^D(P)\right)c_S \\ &\quad +\beta\left(k^S - m^S\left(P(1-\delta) + Ns\left(b - Ns\left(m^S\beta + c_S\right)\right) + E^D(P)\right)\right) \\ &= k^S\beta - E^D(P)\left(c_S + m^S\beta\right) + \frac{1}{2}Ns\left(c_S + m^S\beta\right)\left(-2b + Nsc_S + Nm^S s\beta\right) + \frac{1}{2}b^2 \\ &\quad -P(1-\delta)\left(c_S + m^S\beta\right) \\ &= k^S - m^S P \end{aligned}$$

yielding

$$\begin{aligned} m^S &= c_S\kappa\left(1-\delta\right), \\ k^S(1-\beta) &= -E^D(P)c_S\kappa + \frac{1}{2}Nsc_S\kappa\left(Nsc_S\kappa - 2b\right) + \frac{1}{2}b^2, \\ e^S &= b - Ns\kappa c_S, \\ E^S &= Ns\left(b - Ns\kappa c_S\right), \end{aligned}$$

where, using (1) $e^S = N^2\alpha\kappa\tau s^2 - N\kappa\left(d + N\alpha\tau\right)s + b$ is a convex function of s.

7.2 Equilibrium emissions of non-signatory countries for a fixed s

Assume that $V^D(P; E^S, E^d) = k^D - m^D P$. We then have

$$\begin{aligned} V^D(P; E^S, E^d) &= \max_e \left\{ e\left(b - \frac{e}{2}\right) - \left(P\left(1-\delta\right) + E^S + e + E^d\left(P\right)\right)c_D \right. \\ &\quad \left. +\beta\left(k^D - m^D\left(P\left(1-\delta\right) + E^S + e + E^d\left(P\right)\right)\right)\right\}. \end{aligned}$$

Differentiating w.r.t. emissions yields:

$$\begin{aligned} \frac{d}{de}\left(e\left(b - e/2\right) - \left(P\left(1-\delta\right) + E^S + e + E^d\left(P\right)\right)c_D \right. & \\ \left. +\beta\left(k^D - m^D\left(P\left(1-\delta\right) + E^S + e + E^d\left(P\right)\right)\right)\right) &= b - e - c_D - m^D\beta, \\ \frac{d}{de}\left(b - e - c_D - m^D\beta\right) &= -1, \end{aligned}$$

so that the optimal emissions are given by

$$e^D = b - c_D - m^D\beta,$$

provided the solution is interior. Substituting e^D in (7), we obtain

$$\begin{aligned} V^D(P; E^S, E^d) &= \left(b - c_D - m^D\beta\right)\left(b - \frac{\left(b - c_D - m^D\beta\right)}{2}\right) \\ &\quad - \left(P\left(1-\delta\right) + E^S + \left(b - c_D - m^D\beta\right) + E^d\left(P\right)\right)c_D \\ &\quad +\beta\left(k^D - m^D\left(P\left(1-\delta\right) + E^S + \left(b - c_D - m^D\beta\right) + E^d\left(P\right)\right)\right) \\ &= -\frac{1}{2}\left(c_D + m^D\beta\right)\left(2b - c_D + 2E^S + 2E^d - m^D\beta\right) + k^D\beta + \frac{1}{2}b^2 \\ &\quad -P\left(1-\delta\right)\left(c_D + m^D\beta\right) \\ &= k^D - m^D P, \end{aligned}$$

so that

$$\begin{aligned} m^D &= c_D\kappa(1-\delta), \\ k^D(1-\beta) &= -\frac{1}{2}c_D\kappa\left(2b+2E^S+2E^d(P)-c_D\kappa\right)+\frac{1}{2}b^2, \\ e^D &= b-\kappa c_D, \\ E^D &= \left(N\left(1-s\right)\right)\left(b-\kappa c_D\right), \end{aligned}$$

where, using (1), $e^D = b-\kappa\left(d+\alpha N s\right)$ is a linear decreasing function of s.

7.3 Comparison of emissions of signatories and defectors

Compute

$$\begin{aligned} e^D - e^S &= b-\kappa\left(d+\alpha N s\right)-\left(b-Ns\kappa\left(d+\tau\alpha N\left(1-s\right)\right)\right) \\ &= -\kappa\left(d-Ns\left(d+\alpha\left(N\tau-1\right)\right)+N^2s^2\alpha\tau\right), \end{aligned} \tag{17}$$

which is a concave function of s.

At $s=\frac{1}{N}$, (17) is equal to $\kappa\alpha\left(\tau\left(N-1\right)-1\right)$, which is positive iff $\tau > \frac{1}{N-1}$; assume this condition holds.

At $s=\frac{N-1}{N}$, (17) is equal to

$$\begin{aligned} \kappa\left(d\left(N-2\right)+\alpha\left(\tau-1\right)\left(N-1\right)\right) &> \kappa\left(d\left(N-2\right)+\alpha\left(\frac{1}{N-1}-1\right)\left(N-1\right)\right) \\ &= \kappa\left(N-2\right)\left(d-\alpha\right), \end{aligned}$$

which is positive if $d>\alpha$. Since the function is concave, it cannot be negative in $\left[\frac{1}{N},\frac{N-1}{N}\right]$.

7.4 Equilibrium solution at a fixed s

Total emissions:

$$E(s) = E^S + E^D = N\left(b-\kappa\left(c_D\left(1-s\right)+Ns^2c_S\right)\right)$$

Pollution stock

$$P_t = P_{t-1}\left(1-\delta\right)+N\left(b-\kappa\left(c_D\left(1-s\right)+Ns^2c_S\right)\right)$$

Steady-state

$$P^*(s) = \frac{E(s)}{\delta} = \frac{N\left(b-\kappa\left(c_D\left(1-s\right)+Ns^2c_S\right)\right)}{\delta}$$

Total discounted welfares

$$\begin{aligned} V^S(s,P) &= -c_S\kappa\left(1-\delta\right)P+\frac{2N\kappa c_S\left(-b+\kappa\left(c_D\left(1-s\right)\right)\right)+N^2s^2\kappa c_S^2+b^2}{2(1-\beta)} \\ &= -c_S\kappa\left(1-\delta\right)P-\frac{\kappa c_S}{(1-\beta)}E(s)+\frac{b^2-N^2s^2\kappa^2c_S^2}{2(1-\beta)}, \end{aligned}$$

$$\begin{aligned} V^D(s,P) &= -c_D\kappa(1-\delta)P + \frac{\kappa c_D\left(-2Nb+\kappa\left(2N^2s^2c_S+c_D\left(2N\left(1-s\right)-1\right)\right)\right)+b^2}{2(1-\beta)} \\ &= -c_D\kappa(1-\delta)P - \frac{\kappa c_D}{(1-\beta)}E(s) + \frac{b^2-\kappa^2c_D^2}{2(1-\beta)}. \end{aligned}$$

7.5 Stability condition

At a given P, consider the continuous function

$$w^S(s) = -c_S\kappa\left(1-\delta\right)P + \frac{N\kappa c_S\left(-2b+\kappa\left(2c_D\left(1-s\right)+Ns^2c_S\right)\right)+b^2}{2\left(1-\beta\right)}$$

defined on $\mathbb{R}$. Recalling that c_S and c_D are linear functions of s, it follows that $w^S(s)$ is a polynomial of degree 4 in s. Notice that $w^S(s)$ coincides with the total discounted equilibrium welfare of a signatory country when the pollution stock is P for integer values of $Ns \in [1,N]$.

In the same way, the function

$$w^D(s) = -c_D\kappa(1-\delta)P + \frac{\kappa c_D\left(-2Nb+\kappa\left(2N^2s^2c_S+c_D\left(2N\left(1-s\right)-1\right)\right)\right)+b^2}{2(1-\beta)}$$

is also a polynomial of degree 4 in s, which coincides with the total discounted equilibrium welfare of a non-signatory country when the pollution stock is P for integer values of $Ns \in [0,N-1]$. It follows that the stability function $\Psi(s) = w^S(s) - w^D(\frac{Ns-1}{N})$ is a polynomial function of degree 4 in s which admits at most four real roots. It is straightforward to check that a root $s^0 \in \left[\frac{1}{N},1\right]$ of the stability function satisfies the entry and exit conditions of a coalition $s^* = \left\lfloor Ns^0 \right\rfloor$ if the stability function is decreasing at s^0.

At a given $\bar{P}$, compute

$$\begin{aligned} &w^D(\frac{Ns-1}{N}) = \\ &\quad -\left(c_D-\alpha\right)\kappa(1-\delta)\bar{P} \\ &\quad +\frac{\kappa\left(c_D-\alpha\right)\left(-2Nb+\kappa\left(2\left(Ns-1\right)^2\left(c_S+\alpha\tau\right)+\left(c_D-\alpha\right)\left(2N\left(1-\left(\frac{Ns-1}{N}\right)\right)-1\right)\right)\right)}{2(1-\beta)} \\ &\quad +\frac{b^2}{2(1-\beta)}. \end{aligned}$$

The stability function is equal to 0 when

$$\begin{aligned} &-c_S\kappa\left(1-\delta\right)\bar{P} + \frac{N\kappa c_S\left(-2b+\kappa\left(2c_D\left(1-s\right)+Ns^2c_S\right)\right)+b^2}{2\left(1-\beta\right)} \\ = &-\left(c_D-\alpha\right)\kappa(1-\delta)\bar{P} \\ &+\frac{\kappa\left(c_D-\alpha\right)\left(-2Nb+\kappa\left(2(Ns-1)^2\left(c_S+\alpha\tau\right)+\left(c_D-\alpha\right)\left(2N\left(1-s\right)+1\right)\right)\right)+b^2}{2(1-\beta)}. \end{aligned}$$

Rearranging yields:

$$\begin{aligned} 2\left(1-\beta\right)\left(1-\delta\right)\bar{P}\left(c_D-\alpha-c_S\right) &= -2bN\left(c_D-\alpha-c_S\right) \\ &\quad +2\kappa\left(Ns-1\right)^2\left(c_S+\alpha\tau\right)\left(c_D-\alpha\right) \\ &\quad +\kappa\left(2N\left(1-s\right)+1\right)\left(c_D-\alpha\right)^2 \\ &\quad -N\kappa c_S\left(2c_D\left(1-s\right)+Ns^2c_S\right). \end{aligned}$$

A necessary condition for stability at $\bar{P}$ is therefore:

$$\begin{aligned} \left(1-\delta\right)\left(1-\beta\right)\bar{P} &= -bN+\frac{\kappa\left(Ns-1\right)^2\left(c_S+\alpha\tau\right)\left(c_D-\alpha\right)}{\left(c_D-\alpha-c_S\right)} \\ &\quad +\kappa\frac{\left(2N\left(1-s\right)+1\right)\left(c_D-\alpha\right)^2-Nc_S\left(2c_D\left(1-s\right)+Ns^2c_S\right)}{2\left(c_D-\alpha-c_S\right)}. \end{aligned}$$

7.6 Steady-state of the dynamic system

At the steady-state, if $s^*\in\left[\frac{1}{N},1\right]$ and $P^*=P^*(s^*)$, then

$$V^S(s^*,P^*)=s^*V^S(s^*,P^*)+(1-s^*)V^D(s^*-\frac{1}{N},P^*)$$

or equivalently

$$V^S(s^*,P^*)=V^D(s^*-\frac{1}{N},P^*),$$

which corresponds to a root of the stability function at (s^*,P^*). Now, notice that in the neighborhood of a root in the stability function, the update mechanism moves the proportion of signatories away from the root when the stability function is increasing, and towards the roots when the stability function is decreasing. A coalition size $\lfloor Ns^*\rfloor\in\left[\frac{1}{N},1\right]$ thus satisfies the entry and exit conditions.

An equivalent adjustment mechanism, which is well defined on $\left[0,\frac{N-1}{N}\right]$ is the following

$$s_{t+1}=s_t\frac{V^S(s_t+\frac{1}{N},P_t)}{s_tV^S(s_t+\frac{1}{N},P_t)+(1-s_t)V^D(s_t,P_t)}. \tag{18}$$

One can use (18) whenever the proportion of signatories falls below $1/N$, and revert to (16) if it increases past $\frac{N-1}{N}$, for instance. A steady state of (18) corresponds to a root of the stability function, where the coalition size satisfying the entry and exit conditions is $\lceil Ns^*\rceil\in\left[0,\frac{N-1}{N}\right]$.

7.7 No cost for punishing

In this case, the difference between emissions of defectors and signatories is given by

$$e^D-e^S=\kappa\left(-d+Ns\left(d-\alpha\right)\right).$$

Notice that the function

$$w^S(s) = -d\kappa\left(1-\delta\right)P + \frac{N\kappa d\left(-2b+\kappa\left(2c_D\left(1-s\right)+Ns^2d\right)\right)+b^2}{2\left(1-\beta\right)}$$

is quadratic, while

$$w^D(s) = -c_D\kappa(1-\delta)P + \frac{\kappa c_D\left(-2Nb+\kappa\left(2N^2s^2d+c_D\left(2N\left(1-s\right)-1\right)\right)\right)+b^2}{2(1-\beta)}$$

is cubic. It follows that the stability function $\Psi(\frac{Ns+1}{N}) = w^S(\frac{Ns+1}{N}) - w^D(s)$ is a polynomial function of degree 3 in s.

The coefficient of s^3 in the stability function is $N^3\alpha\kappa^2\frac{\alpha-d}{1-\beta}$, and it has the same sign as $\alpha - d$. At $s = 0$, compute

$$\Psi(\frac{1}{N}) = d\alpha\frac{\kappa^2}{1-\beta}\left(N-1\right) > 0$$

and

$$\Psi'(\frac{1}{N}) = N\frac{\kappa}{1-\beta}\left(P\alpha\left(1-\delta\right)\left(1-\beta\right)+d\kappa\left(d-\alpha\right)+N\alpha\left(b-d\kappa\right)\right).$$

i Under the assumption that $b > \kappa\left(d+\alpha Ns\right)$ (interior solution), than if $d > \alpha$, the stability function is increasing and positive at $s = 0$. Since the coefficient of s^3 is negative, there is at most one root of the stability function where it is decreasing which is greater than 0; since a stable coalition is an integer value not smaller than that root, $s = 0$ is not a stable coalition.

ii If on the other hand $d < \alpha$, the coefficient of s^3 is positive, and the stability function is positive at $s = 0$. There is at most one root of the stability function where it is decreasing which is greater than 0, and $s = 0$ is not a stable coalition.

iii If $d = \alpha$, the stability function is quadratic, concave, and positive at $s = 0$. There is at most one root of the stability function where it is decreasing which is greater than 0, and $s = 0$ is not a stable coalition.

7.8 Endogenous punishment

The smallest level of α for which full cooperation is achieved at the steady state of pollution satisfies (15) at $\bar{P} = N\left(\frac{b-Nd\kappa}{\delta}\right)$ and $s = 1$, or equivalently solves

$$\begin{aligned}
2\left(1-\beta\right)\left(1-\delta\right)N\left(\frac{b-Nd\kappa}{\delta}\right)\left(\alpha\left(N-1\right)\right) &= -2bN\left(\alpha\left(N-1\right)\right) \\
&\quad +2\kappa\left(d+\alpha\tau\right)\left(N-1\right)^2\left(d+\alpha\left(N-1\right)\right) \\
&\quad +\kappa\left(d+\alpha\left(N-1\right)\right)^2 \\
&\quad -N^2d^2\kappa,
\end{aligned}$$

which reduces to

$$\begin{aligned} & \alpha^2\kappa^2\delta\left(N-1\right)\left(2\tau\left(N-1\right)+1\right) \\ +2\alpha & \left(-bN+d\kappa\left(\kappa\delta\left(\tau-2\right)\left(N-1\right)+N^2\right)\right) \\ & +d^2\kappa^2\delta\left(N-3\right) \quad = \quad 0. \end{aligned}$$