FACULTY OF ECONOMICS AND
APPLIED ECONOMIC SCIENCES
CENTER FOR ECONOMIC STUDIES
ENERGY, TRANSPORT & ENVIRONMENT

KATHOLIEKE
UNIVERSITEIT
LEUVEN

WORKING PAPER SERIES
n°2001-03

On the farsighted stability of the Kyoto Protocol

Johan Eyckmans (K.U.Leuven)

January 2001

contact:
Isabelle Benoit
KULeuven-CES
Naamsestraat 69, B-3000 Leuven (Belgium)
tel: +32 (0) 16 32.66.33
fax: +32 (0) 16 32.69.10
e-mail: Isabelle.Benoit@econ.kuleuven.ac.be
http://www.econ.kuleuven.ac.be/ew/academic/energmil

On the farsighted stability of the Kyoto Protocol [*]

Johan Eyckmans[†]

January 25, 2001

Abstract

This paper investigates the coalitional stability of the 1997 Kyoto Protocol on the emissions of greenhouse gases. Unlike conventional coalition stability tests we assume that potential deviators are farsighted in the sense of Chwe (1994) and take into account possible subsequent deviations by the remaining players. In the empirical part of the paper, a Partial Agreement Nash Equilibrium w.r.t. to the Kyoto coalition is computed with a stylized dynamic integrated assessment model that resembles closely the RICE model by Nordhaus and Yang (1996). The simulations show that the Kyoto coalition is more stable than suggested by conventional myopic stability concepts but that the stability analysis is very sensitive to the coalitional surplus sharing rule.

Keywords

climate change, coalition stability, farsightedness,

JEL codes:

C72, D62, Q2

[*] The author is grateful to the Bren School of Environmental Sciences and Management at the University of California at Santa Barbara for their hospitality during the fall semester 1999. Stimulating suggestions by participants of the 5th Coalition Formation Workshop in Barcelona and seminar participants at the Fondazione ENI Enrico Mattei in Venice are gratefully acknowledged. This work is part of the CLIMNEG project funded by OSTC, the Belgian Federal Office for Scientific, Technical and Scientific Cooperation, contract CG/DD/243.

[†] Postdoctoral Research Assistant of the Belgian National Fund for Scientific Research, FWO Vlaanderen. Katholieke Universiteit Leuven, CES, Naamsestraat 69, B-3000 Leuven, Belgium, e-mail: johan.eyckmans@econ.kuleuven.ac.be

Contents

1 Introduction and summary of results

In negotiations on international environmental agreements, we often observe that cooperation is only partial, i.e. only a subset of the countries involved in the pollution problem actually agrees upon pollution control measures. The 1997 Kyoto Protocol on the emissions of greenhouse gases (GHG) is a prominent example. The Kyoto Protocol requires that the joint emissions of greenhouse gases of all Annex B countries[1] should be lower by approximately 5 % compared to their 1990 emission levels by the compliance period 2008-2012. Different emission reduction targets were assigned to the Annex B countries but at the same time the Protocol provides for flexible mechanisms (e.g. emission trading) to redistribute these obligations and to safeguard the overall reduction target of minus 5 %.

The big issue is of course whether this Protocol is stable, i.e. whether the Annex B countries have an incentive to keep to their promises. Since the greenhouse problem is a perfect example of a public bad, one might presume that strong free riding incentives are present. Individual Annex B countries might be better off leaving the agreement and hoping that the remaining countries keep to their promises to reduce GHG emissions. The deviator would enjoy the benefits of reduced greenhouse warming without sharing in the costs. Several preliminary indications do indeed cast doubts on the stability of the Kyoto Protocol. First, most Annex B countries' emissions of greenhouse gases continue to rise at a non checked rate. Most signatories seem to postpone GHG emission reduction efforts until the very last moment. Secondly, most experts agree that it is unlikely that the United States Senate will approve the ratification of the Kyoto Protocol. If the US would not ratify the Protocol, stability of the agreement is severely jeopardised. Finally, the failure to achieve an agreement during the last Conference of the Parties (COP 7) in The Hague painfully revealed the deep differences in opinion between the US and the EU concerning the interpretation of the Kyoto Protocol.

In the context of a stylized theoretical model with symmetrical players, the stability of international environmental agreements has been studied extensively, see among others Barrett (1994,1997), Carraro and Siniscalco (1993) and Carraro (1999) and the references therein. Most of this literature uses the coalition stability notion introduced by d'Aspremont et. al (1983) in the context of cartel formation. According to this notion of stability, a coalition is said to be stable if (*i*) none of its members has a profitable individual deviation, and (*ii*) none of the non members wants to join the coalition. Two typical results emerge from the literature on voluntary environmental agreements. Firstly, if the stakes are high (i.e. if both the potential environmental damage and the cost to reduce emissions are high),

[1] The countries that agreed to the 1997 Kyoto Protocol are listed in Annex B of the Protocol text. Ever since, the signatories are called Annex B countries. Roughly speaking, the Annex B consists of the OECD countries, the former Soviet Union and Eastern European countries.

the grand coalition is unlikely to be stable. If there exists a stable coalition at all, it will probably be a small one in terms of the number of signatories. Secondly, if we observe an international environmental agreement involving many participants, it often produces little additional environmental protection over the laissez-faire situation, see Barrett (1997) and Carraro (1999).

However, it has been argued by, among others, Chwe (1994) that this notion of stability is myopic since it assumes that a country with a profitable deviation strategy will use this strategy, no matter the consequences, i.e. no matter possible further deviations by the remaining players in the agreement. Consider the example of the US and assume the US Senate rejects the ratification of the Kyoto Protocol. Will the US be able to reap the benefits of this free riding strategy? Clearly, the answer depends on the behaviour of the remaining Annex B countries. If they stick to their Kyoto obligations, the US has probably a credible deviating strategy. It can the reap the benefits of emission reduction policies without contributing itself to it. If, on the other hand, some of the other Annex B countries decide to leave the agreement, the overall level of emission abatement will be lower. In the extreme case, the initial deviation of the US might trigger so many deviations by other Annex B countries that the Protocol cannot come into effect because the required emission and ratification quota[2] cannot be achieved. We might end up in a situation of complete absence of cooperation which might be undesirable for the US as well because it would suffer from important climate change damages.

In order to simulate the effects of the Kyoto Protocol and of possible deviations by its signatories, the following issues need to be addressed: (1) Which equilibrium concept do we use to describe this situation of partial cooperation? What assumption do we make concerning the behaviour of outsiders to the Protocol? Do they try to hurt the coalition or do they act to help the coalition? (2) How do the members of a coalition divide the surplus of cooperation among each other? Do they make monetary transfers, do they set up a system of tradeable emission permits among themselves and how do they allocate the permits? (3) If an Annex B country considers deviating, how does it evaluate the possible further deviations by the remaining countries?

This paper makes specific but reasonable assumptions on these three issues and investigates what they imply in the standard RICE simulation model describing the world greenhouse economy. The simulations suggests the following. (1) The surplus sharing rule within a coalition is very important to the final coalition structure that can form. In particular, allowing for emission trading and grand fathering initial permits seems to provide better opportunities for stability than no transfers at all. (2) Conventional myopic stability analysis suggests that several signatories of the Kyoto Protocol would have a profitable free riding

[2] In order to come into effect, the Protocol has to be ratified by at least 50% of the signatories representing at least 50% of GHG emissions of the Annex B countries.

strategy. (3) Introducing farsightedness strongly restricts the number of credible free riding strategies. In particular, only the Former Soviet Union seems to have a potentially credible deviation strategy in the absence of transfers. If emission trading is allowed for, only Japan has a credible deviation. (4) With emission trading based on the Kyoto Protocol emission assignments, the coalition defined by the Annex B countries minus Japan is stable in the farsighted sense.

This paper is organised as follows. Section 2 describes a standard optimal growth model with a climate externality. The formulation closely resembles the formulation of the RICE model by Nordhaus and Yang (1996). Some simplifying assumptions were made to ensure tractability of the model. The concept of a Partial Agreement Equilibrium (PANE) for a coalition is defined in section 3. This equilibrium concept was introduced by Chander and Tulkens (1995) and can be described as a Nash equilibrium where the players are coalitions instead of individual countries. Section 4 describes the stability concept that will be used and introduces the notion of farsightedness. This notion as based upon the largest consistent set in Chwe (1994). While considering a deviation, a country is interested in the final outcome, i.e. it takes into account all further possible deviations that might occur. Section 5 reports the simulation results with the RICE model. Section 6 concludes.

2 Model description

The model we will use for defining partial cooperation in greenhouse negotiations resembles closely the integrated assessment model RICE introduced by Nordhaus and Yang (1996). We introduced some simplifications in order to enhance the tractability of the theoretical model. For the simulations however we resort to the more elaborate specification of the RICE model. We start by describing the theoretical model.

Consider an optimal growth model without international trade. We will use a discrete time model with a finite horizon. N denotes the set of regions[3] indexed $i = 1, 2, \dots, n$. Growth is assumed to be driven by exogenous population growth and technological change and by endogenous capital accumulation. The following equations describe the *economy* of a country i at time t:

$$Y_{i,t} \geq Z_{i,t} + I_{i,t} + C_i(\mu_{i,t}) + D_i(\Delta T_t) \tag{1}$$

$$Y_{i,t} = A_{i,t} F_i(K_{i,t}) \tag{2}$$

$$K_{i,t+1} = [1 - \delta_K] K_{i,t} + I_{i,t} \qquad \text{with} \quad K_{i,0} \text{ given} \tag{3}$$

[3] In the sequel we will always speak of "regions" even if a region contains only one country.

A complete list of all variables and parameters is given in appendix. Equation (1) is a standard budget equation requiring that in every period production $Y_{i,t}$ is sufficient to cover the claims of consumption $Z_{i,t}$, investment $I_{i,t}$, cost of abatement $C_i(\mu_{i,t})$ and climate change damage $D_i(\Delta T_t)$ upon production. The costs of abatement and of climate change damage functions are both assumed strictly increasing and strictly convex in abatement $\mu_{i,t}$ and temperature change ΔT_t respectively[4]. (2) defines production as a strictly increasing and strictly concave function of capital $K_{i,t}$ input. $A_{i,t}$ measures overall productivity. It is assumed that productivity increases exogenously as time goes by and technological progress is Hicks neutral. Since labour supply is assumed exogenous, this argument is omitted in the production function. Labour input is subsumed in the functional form of the productivity measure $A_{i,t}$. Finally, expression (3) is a standard capital accumulation equation where δ_K stands for the rate of capital depreciation.

This model of the world economy is coupled to a simple model of global mean temperature change. The *carbon emissions, the carbon cycle and climate module* are respectively modelled by the following three equations:

$$E_{i,t} = \sigma_{i,t}\,[1 - \mu_{i,t}]\,Y_{i,t} \tag{4}$$

$$M_{t+1} = [1 - \delta_M]\,M_t + \beta \sum_{i \in N} E_{i,t} \qquad \text{with } M_0 \text{ given} \tag{5}$$

$$\Delta T_t = G(M_t) \tag{6}$$

According to expression (4), carbon emissions are proportional to production. The emissions to output ratio $\sigma_{i,t}$ declines exogenously over time due to an assumed autonomous energy efficiency increase (AEEI). Emissions can be reduced at a rate $\mu_{i,t} \in [0,1]$ in every period though this is costly according to equation (1). Equation (5) describes the accumulation of carbon in the atmosphere. This process is modelled similarly to a standard capital accumulation process where δ_M denotes the natural decay rate of atmospheric carbon concentrations and β is the airborne fraction of carbon emissions. Expression (6) translates atmospheric

[4] This formulation is different from the one used by Nordhaus and Yang (1996) because we use an additive instead of a multiplicative formulation of climate change damages. Translated into our notation, Nordhaus and Yang's (1996) formulation of the budget equation (1) would be given by:

$$\Omega_{i,t}\,Y_{i,t} \equiv \frac{1 - C_{i,t}/Y_{i,t}}{1 + D_{i,t}/Y_{i,t}}\,Y_{i,t} = Z_{i,t} + I_{i,t}$$

Conceptually, both formulations are identical in the sense that the costs of emission abatement and of damage from climate change reduce the amount of production that can be devoted to consumption or investment. The difference between both formulation stems from the fact that Nordhaus and Yang (1996) allow for cross effects between emission abatement costs and climate change damages. This type of cross effects are precluded by our formulation.

carbon concentration levels into global mean temperature change. We assume that G is a continuous differentiable and increasing function. The function G can also stand for a more complex relationship between atmospheric carbon concentration and temperature change as is the case in the RICE model.

It is assumed that countries are choosing consumption, investment and emission paths that maximize their lifetime discounted consumption. Lifetime utility of player is denoted by W_i:

$$W_i \; = \sum_{t=0}^{T} \frac{Z_{i,t}}{[1+\rho_i]^t} \; + \; w_i(K_{i,T+1}) \tag{7}$$

where ρ_i stands for the discount rate used by country i. The strictly increasing and strictly concave function w_i stands for the scrap value of the terminal capital stock. Notice that in contrast to Nordhaus and Yang (1996) utility is simply linear in consumption. We make this simplification in order to represent the global carbon emission game as a transferable utility (TU) game.

3 Partial Agreement Nash Equilibrium w.r.t. a coalition

3.1 Definition

In reality, we often observe partial or intermediate cooperation in international environmental agreements. Hence, only some subgroup of countries affected by the problem agrees to coordinate its emission reduction policies. The 1997 Kyoto protocol is a prominent example of partial cooperation. In order to characterize this situation of partial cooperation, I will use the concept of a *Partial Agreement Nash Equilibrium w.r.t. a coalition* (PANE in the sequel). This equilibrium concept was introduced by Chander and Tulkens (1995) and (1997) in the context of a static model but it can readily be extended to a dynamic framework.

Suppose a coalition $S \subseteq N$ forms. In a PANE w.r.t. coalition S, the coalition S chooses actions that are most beneficial from the group point of view while the outsiders to the coalition choose actions that maximize their individual utility. The PANE w.r.t. coalition S can be interpreted as a Nash equilibrium with player set $(S, \{j\}_{j \in N \setminus S})$. If one reinterprets the game such that the coalition of cooperating countries stands for only one player, the PANE w.r.t. a coalition in the original game is equivalent to an ordinary Nash equilibrium in the new game. The coalition S coordinates its policies taking as given the emission strategies of the outsiders who, on their turn, are playing a non cooperative Nash strategy against S.

Definition 1 *A Partial Agreement Nash Equilibrium (PANE) w.r.t. coalition $S \subseteq N$ is a combination of strategies $(Z,\mu) \in I\!R_+^{2nt}$ that solves simultaneously the following maximization problems:*

a. for all insiders $j \in S$*:*

$$\max_{Z_{j,t},\,\mu_{j,t}} \quad \sum_{t=0}^{T} \sum_{j \in S} \frac{Z_{j,t}}{[1+\rho_j]^t} \;+\; w_j(K_{j,T+1}) \tag{8}$$

subject to (1), (2), (3), (4), (5) and (6);

b. for all outsiders $i \in N \setminus S$*:*

$$\max_{Z_{i,t},\,\mu_{i,t}} \quad \sum_{t=0}^{T} \frac{Z_{i,t}}{[1+\rho_i]^t} \;+\; w_i(K_{i,T+1}) \tag{9}$$

subject to (1), (2), (3), (4), (5) and (6).

We will denote lifetime utility of a player i under the PANE w.r.t. coalition S by W_i^S. The PANE definition encompasses both the definition of Pareto efficient allocations for $S = N$ and the definition of an ordinary Nash equilibrium for $S = \{i\}$. When the players are not cooperating and are organized in singletons, we will call this coalition structure the *trivial coalition structure*. We will assume that for any possible coalition $S \subseteq N$, there always exists a PANE w.r.t. S and that it is unique. Sufficient continuity and concavity assumptions can be found for this result to hold.

3.2 First-order conditions for outsiders

Consider the utility maximization problem faced by an outsider $i \in N \setminus S$:

$$\max_{Z_{i,t},\,I_{i,t},\,K_{i,t},\,\mu_{i,t},\,M_t} \quad \sum_{t=0}^{T} \frac{Z_{i,t}}{[1+\rho_i]^t} \;+\; w_i(K_{i,T+1}) \tag{10}$$

subject to (for all $0 \leq t \leq T$):

$$A_{i,t}\,F_i(K_{i,t}) \geq Z_{i,t} \;+\; I_{i,t} \;+\; C_i(\mu_{i,t}) \;+\; D_i(G(M_t)) \qquad [\zeta_{i,t}]$$

$$K_{i,t+1} = [1 \;-\; \delta_K]\,K_{i,t} \;+\; I_{i,t} \qquad [\psi_{i,t}]$$

$$M_{t+1} = [1 \;-\; \delta_M]\,M_t \;+\; \beta\,\sigma_i\,[1 \;-\; \mu_{i,t}]\,A_{i,t}\,F_i(K_{i,t}) \;+\; \beta \sum_{j \neq i} E_{j,t}^S \qquad [\phi_{i,t}]$$

with M_0 and $K_{i,0}$ given and nonnegativity restrictions for $Z_{i,t}, I_{i,t}, K_{i,t}, \mu_{i,t}, M_t$. We associate Lagrange multipliers $\zeta_{i,t}$ to the resource constraint, $\psi_{i,t}$ to the capital accumulation constraint and $\phi_{i,t}$ to the carbon accumulation process. First-order conditions for all $0 \leq t \leq T$ for an interior optimum are given by (the superscript "S" refers to the equilibrium values of the variables for the PANE w.r.t. coalition S):

$$\zeta_{i,t}^S = \frac{1}{[1+\rho_i]^t} = \psi_{i,t}^S \tag{11}$$

$$\begin{aligned}\psi_{i,t-1}^S &= \psi_{i,t}^S \, [A_{i,t} \, F_i'(K_{i,t}^S) \, + \, [1-\delta_K]] \\ &\qquad - \beta \, \sigma_{i,t} \, [1-\mu_{i,t}^S] \, A_{i,t} \, F_i'(K_{i,t}^S) \, \phi_{i,t}^S\end{aligned} \tag{12}$$

$$\psi_{i,T}^S = w_i'(K_{i,T+1}^S) \tag{13}$$

$$\psi_{i,t}^S \, C_i'(\mu_{i,t}^S) = \beta \, \sigma_{i,t} \, A_{i,t} \, F_i(K_{i,t}^S) \, \phi_{i,t}^S \tag{14}$$

$$\phi_{i,t-1}^S = G'(M_t^S) \, \psi_{i,t}^S \, D_i'(G(M_t^S)) \, + \, [1 \, - \, \delta_M] \, \phi_{i,t}^S \qquad \phi_{i,T}^S = 0 \tag{15}$$

The first condition (11) say that the shadow cost of capital equals the shadow cost of the resource constraint and that both are equal to the discount factor. The evolution of the capital stock is described by conditions (12). The terminal capital stock is determined by (13). (14) determines the optimal amount of carbon emission control for country i. Expression (15) describes the evolution of the shadow price of atmospheric carbon concentration. In the sequel, the shadow price to country i of carbon accumulation in the atmosphere will be referred to as the *carbon tax for country i.* In the last period, this shadow price is zero because there is no valuation of the terminal carbon concentration.

We start by solving the difference equation (15). From the terminal condition $\phi_{i,T}^S = 0$ and by solving iteratively from (15), it can be shown that the carbon tax for an outsider at any period t is equal to the sum of future marginal damage caused by an additional unit of carbon emissions at time t, evaluated at the appropriate discount factor:

$$\phi_{i,t}^S = \left[\frac{1}{1+\rho_i}\right]^{t+1} \sum_{\tau=t+1}^{T} \left[\frac{1-\delta_M}{1+\rho_i}\right]^{\tau-t-1} G'(M_\tau^S) \, D_i'(G(M_\tau^S)) \tag{16}$$

Notice that the carbon tax for country i only takes into account the climate change damage occurring within its territory, spill over effects to neighbouring countries are not taken into account in country i's individual decision process. Substituting for the carbon tax in (14), we derive the rule driving the optimal amount of carbon emission control for an outsider country. In particular, in a PANE, every outsider country equalizes its marginal costs of abatement (per ton of carbon) to the marginal damage from the resulting climate change (all quantities are evaluated at the appropriate discount factor):

$$\frac{C_i'(\mu_{i,t}^S)}{\sigma_{i,t}\,A_{i,t}\,F_i(K_{i,t}^S)} \;=\; \frac{\beta\,\phi_{i,t}^S}{\zeta_{i,t}^S} \;=\; \frac{\beta}{1+\rho_i}\sum_{\tau=t+1}^{T}\left[\frac{1-\delta_M}{1+\rho_i}\right]^{\tau-t-1} G'(M_\tau^S)\,D_i'(G(M_\tau^S)) \tag{17}$$

This is the traditional optimality condition for a noncooperative Nash behaviour saying that individual marginal costs should be equal to individual marginal benefits of taking effort. The left hand side (LHS) of the expression stands for the marginal cost for region i of reducing its carbon emissions by an additional ton in period t. The denominator denotes gross emissions without abatement and is used to convert the units of the marginal abatement costs into US$ per ton of carbon[5]. The RHS of the expression consists of the sum of country i's discounted future marginal damages from climate change. Only the fraction of the emissions that become actually airborne is taken into account (multiplication by β). The term $[(1-\delta_M)/(1+\rho_i)]^{\tau-t-1}$ is a deflation effect for the valuation of marginal damage in period t. Because of the discount rate and natural decay rate of carbon concentrations in the atmosphere, the effect of emitting one extra ton of carbon at time t gradually dies off.

We now turn to the condition that drives capital accumulation for country i. Substituting (11) and (14) into (12), the latter condition can be written as follows:

$$\rho_i \;+\; \delta_K \;=\; A_{i,t}\,F_i'(K_{i,t}^S)\left[1 \;-\; \frac{[1-\mu_{i,t}^S]\,C_i'(\mu_{i,t}^S)}{A_{i,t}\,F_i(K_{i,t}^S)}\right] \tag{18}$$

This condition is the translation of the *Ramsey-Keynes optimal consumption/investment rule* for an optimal growth model. It says that the along the optimal investment path, marginal product of capital should equal the sum of the rate of time preference and capital depreciation. Notice that because of the climate externality, the marginal product of capital is corrected to take into account the marginal damage from increased emissions as production is expanded. However, outsiders only take into account damages on their territory, they do not care about negative climate change externalities they inflict upon their neighbouring countries. If country i would not value climate change damages, $D_i'(\Delta T_t) = 0$ for all ΔT_t, then one sees from condition (17) that it would not reduce its emissions, $\mu_{i,t} = 0$. This implies that the Ramsey-Keynes rule boils down to its familiar form $\rho_i + \delta_K = A_{i,t}\,F_i'(K_{i,t}^S)$.

3.3 First-order conditions for insiders

The members of coalition S are assumed to coordinate their investment and emission strategies. Since utility is assumed quasi-linear, the (restricted) Pareto efficient allocation for coalition

[5] Recall that $\mu_{i,t} \in [0,1]$ has no dimension since it is the fraction of emissions that are abated.

S can be found by maximizing an unweighted sum of the members' utilities:

$$\max_{Z_{i,t},\, I_{i,t},\, K_{i,t},\, \mu_{i,t},\, M_t} \quad \sum_{t=0}^{T} \sum_{j \in S} \frac{Z_{j,t}}{[1+\rho]^t} \; + \; w_j(K_{j,T+1}) \tag{19}$$

subject to (for all $0 \leq t \leq T$):

$$A_{j,t}\, F_i(K_{j,t}) \geq Z_{j,t} \; + \; I_{j,t} \; + \; C_j(\mu_{j,t}) \; + \; D_j(G(M_t)) \qquad \forall j \in S \qquad [\zeta_{j,t}]$$

$$K_{j,t+1} = [1 \; - \; \delta_K]\, K_{j,t} \; + \; I_{j,t} \qquad \forall j \in S \qquad [\psi_{j,t}]$$

$$M_{t+1} = [1 \; - \; \delta_M]\, M_t \; + \; \beta \sum_{j \in S} \sigma_j [1 \; - \; \mu_{j,t}] A_{j,t} F_j(K_{j,t}) \; + \; \beta \sum_{j \in N \setminus S} E^S_{j,t} \qquad [\phi_{i,t}]$$

We associate Lagrange multipliers $\zeta_{j,t}$ to the resource constraints, $\psi_{j,t}$ to the individual capital accumulation constraints and ϕ_t to the carbon accumulation process. First-order conditions for all $j \in S$ and $0 \leq t \leq T$ for an interior optimum are given by (the "S" superscript refers to the values of the variables at the PANE solution):

$$\zeta^S_{i,t} \; = \; \frac{1}{[1+\rho_i]^t} \; = \; \psi^S_{i,t} \tag{20}$$

$$\begin{aligned} \psi^S_{i,t-1} \; = \; & \psi^S_{i,t} \; [A_{i,t}\, F_i'(K^S_{i,t}) \; + \; [1 - \delta_K]] \\ & - \beta\, \sigma_{i,t}\, [1 - \mu^S_{i,t}]\, A_{i,t}\, F_i'(K^S_{i,t})\, \phi^S_t \end{aligned} \tag{21}$$

$$\psi^S_{i,T} \; = \; w_i'(K^S_{i,T+1}) \tag{22}$$

$$\psi^S_{i,t}\, C_i'(\mu^S_{i,t}) \; = \; \beta\, \sigma_{i,t}\, A_{i,t}\, F_i(K^S_{i,t})\, \phi^S_t \tag{23}$$

$$\phi^S_{t-1} \; = \; G'(M^S_t) \sum_{j \in S} \psi^S_{j,t}\, D_j'(G(M^S_t)) \; + \; [1 \; - \; \delta_M]\, \phi^S_t, \quad \phi^S_T = 0 \tag{24}$$

According to (21), the shadow cost of the resource constraint equals the shadow cost of capital. The terminal capital stock is described by condition (22). Last period's shadow price of capital equals the marginal valuation of the terminal capital stock. (23) determines the optimal amount of carbon emission control for country i. In a PANE w.r.t. coalition S, the marginal abatement costs are a function of the shadow cost of atmospheric carbon concentrations. Expression (24) describes the evolution of the shadow price of atmospheric carbon concentration.

From the terminal condition $\phi^S_T = 0$, it follows from (24) through iterative substitution that the carbon tax at any period t is equal to the weighted sum of all future discounted marginal damages experienced by all insiders in S:

$$\phi_t^S \;=\; \sum_{\tau=t+1}^{T} [1-\delta_M]^{\tau-t-1}\, G'(M_\tau^S) \sum_{j\in S} \frac{D_j'(G(M_\tau^S))}{[1+\rho_j]^\tau} \tag{25}$$

Notice that the optimal carbon tax takes into account the climate change damage affecting all coalition members but it does not consider the spill over effects to the outsiders. Substituting for the carbon tax in (23), we can derive the rule driving the optimal amount of carbon emission control for an insider country i in period t:

$$\frac{C_i'(\mu_{i,t}^S)}{\sigma_{i,t}\, A_{i,t}\, F_i(K_{i,t}^S)} \;=\; \beta\,[1+\rho_i]^t \sum_{\tau=t+1}^{T} G'(M_\tau^S)\,[1-\delta_M]^{\tau-t-1} \sum_{j\in S} \frac{D_j'(G(M_\tau^S))}{[1+\rho_j]^\tau} \tag{26}$$

This rule will be referred to in the sequel as the *restricted Samuelson rule for the optimal emission reductions by coalition S*. It is a dynamic extension of the traditional optimality rule for static public good models that was first stated by Samuelson (1954). The left hand side (LHS) of the expression stands for the marginal cost for region i of reducing its carbon emissions by an additional ton in period t. The RHS of the expression consists of the sum of all insiders' discounted future marginal damages from climate change, multiplied by the inverse of the discount factor. If all insiders would be characterized by the same discount rate ($\rho_i = \rho$, $\forall i \in S$, the Samuelson rule (26) thus says that all regions should reduce their emissions in such a way that their marginal abatement costs in each period t be equalized. Hence, it induces both cost efficiency and allocative efficiency for the insider's coalition. For $S = N$, the restricted Samuelson rule recovers the traditional Samuelson rule which would internalize marginal damages for all countries in the world. If discount rates differ across insiders, the Samuelson rule does not induce cost efficiency since marginal abatement costs are inversely proportional to the discount factor $1/[1+\rho_i]^t$. Countries with relatively high discount rates ρ_i will have to abate relatively more.

We now derive the condition for the optimal accumulation of capital in the presence of an environmental externality. Substituting (20) into (22), we obtain:

$$\rho_i \,+\, \delta_K \;=\; A_{i,t}\, F_i'(K_{i,t}^S)\, \left[1 \;-\; \frac{[1-\mu_{i,t}^S]\, C_i'(\mu_{i,t}^S)}{A_{i,t}\, F_i(K_{i,t}^S)}\right] \tag{27}$$

As for the outsiders, this investment rule will be referred to as the *Ramsey-Keynes* optimal investment rule. Though the rule looks exactly the same for outsiders and insiders, the difference between both stems from what climate change damages are internalized, For insiders, marginal costs reflect all marginal damages inflicted upon other coalition members. For outsiders, it only reflects domestic damages.

3.4 Surplus sharing rules for the insiders

In the definition of a PANE w.r.t. coalition S, we did not allow for transfers of consumption among the insiders. However, in many international environmental agreements, some form of transfers is adopted, sometimes in the form of emission trading. Transfers and emission trading can easily be incorporated in our model by rewriting the resource constraints for insiders of coalition S as follows:

$$A_{i,t}F_i(K_{i,t}) \;=\; Z_{i,t} + I_{i,t} + C_i(\mu_{i,t}) + D_i(G(M_t)) + T_{i,t} \qquad \forall i \in S \tag{28}$$

where $T_{i,t}$ denotes the transfer received (or paid) by country i in period t. Transfers to outsiders are not allowed for and we require that transfers sum to zero in every period: $\sum_{j\in S} T_{j,t} = 0$. Since utility is linear in consumption, there are no income effects from redistributing consumption. Therefore, the allocation of abatement efforts governed by the restricted Samuelson rule (26) does not change if we reshuffle consumption through transfers. In other words still, the solution to the insiders' optimization problem is undetermined. There are infinitely many ways for an allocation of abatement effort and consumption to satisfy the restricted Samuelson rule, the resource constraints (28) and the restriction $\sum_{j\in S} T_{j,t} = 0$. We will consider two candidates below. In the simulations we will compare both rules in order to determine their impact on the stability of a coalition.

No transfers
An obvious first candidate surplus sharing rule is to give no transfers at all. Hence, every country abates its emissions as prescribed by the restricted Samuelson rule (26) and $T_{i,t} = 0$ for all $i \in S$ and $0 \leq t \leq T$.

Emission trading with grand fathering of permits
A second candidate rule is to consider some kind of emission trading scheme. Define $\bar{E}^S_{i,t}$ as country i's assignment of emission permits and p^S_t as the price.

$$T^S_{i,t} \;=\; p^S_t\,[\bar{E}_{i,t} \;-\; E^S_{i,t}]$$

This rephrases the consumption allocation question as a problem of choosing an initial permit assignment rule. In other words, how do we determine $\bar{E}_{i,t}$? A widely used permit allocation rule is to "grandfather", i.e. give for free, emission allowances in function of baseline emissions. In the model we study, this would amount to saying that (S denotes the trading coalition): $\bar{E}^0_{i,t} = w^S_{i,t} \sum_{j\in S} E^S_{j,t}$ with weights determined by the pre-abatement emissions, or $w^S_{i,t} = \sigma_{i,t} Y^S_{i,t} / \sum_{j\in S} \sigma_{j,t} Y^S_{j,t}$. Though the details of the flexible mechanisms were left open for negotiations during subsequent Conferences Of the Parties and have not been settled yet, one might argue that the Protocol does not envisage this type of grand fathering but rather a division of the permits based upon the Kyoto emission assignments. Hence, the weights are calculated as:

$$w_{i,t} \; = \; \frac{[1 \; - \; \mu_{i,2010}]\, E_{i,2010}}{\sum_{j \in S}[1 \; - \; \mu_{j,2010}]\, E_{j,2010}} \qquad 0 \leq t \leq T \quad \forall i \in S$$

with $E_{i,2010}$ baseline emissions (without abatement) in 2010 and $\mu_{i,2010}$ the emission abatement objectives agreed upon in the 1997 Kyoto Protocol. These objectives are listed in Table 10. Since the Kyoto Protocol does not specify emission abatement efforts $\mu_{i,t}$ beyond the compliance period of 2008-2012, we assume that the weights remain constant over time. It should be mentioned that there are some indications that some Annex B countries want to restrict the trading by means of "caps", i.e, a restriction on how much emission abatement countries would be allowed to buy abroad. We do not consider caps on trading nor transaction or search costs. Therefore, the emission trading scheme in this paper should be interpreted as some kind of an ideal trading that exploits all trading possibilities.

4 Stability of coalitions and farsightedness

4.1 Two extreme views on coalitional stability

Stability in the sense of the core

The question of stability of voluntary international environmental agreements has preoccupied economists for a long time and there have been suggested several concepts of stability. We mention two concepts which can be interpreted as two extremes. First, Chander and Tulkens (1995) define stability by means of the classic core concept. In particular, they define the concept of the γ-*core* using the Partial Agreement Nash Equilibrium w.r.t. a coalition.

Definition 2 (γ-core) *An allocation is said to belong the the* γ-core *if there does not exist a coalition $S \subseteq N$ and a corresponding PANE such that all members of S are at least as well off (and at least one strictly better off) under the alternative allocation compared to the core allocation.*

Allocations in the core always correspond to complete cooperation by the grand coalition N. The γ-core makes use of the PANE concept to capture the behaviour of outsiders to a coalition. In a PANE w.r.t. a coalition S, the outsiders to S are assumed to behave individually and noncooperatively. They play a Nash strategy against S and against all other outsiders. A γ-core allocation is stable in the sense that no coalition can propose a deviation, i.e. a PANE w.r.t. itself, that is profitable for all of its members. For deviations by singletons, this means that the welfare a singleton can achieve under its corresponding PANE is not higher than what it gets at the core allocation. In our notation, if $\bar{W}_i$ denotes the welfare level of player i at a γ-core allocation, it must hold that $\bar{W}_i \geq W_i^{\{i\}} = W_i^{NASH}$ for all $i \in N$. Basically, the coalitional stability idea behind the γ-core assumes that if a single player deviates from the grand coalition N, this will lead to a complete disintegration of the coalition such that we end up in the noncooperative Nash equilibrium in which all players act as singletons.

Stable coalitions
Another line of literature in environmental economics, see among others Barrett (1994) and Carraro and Siniscalco (1993), has focussed upon the stability concept introduced by d'Aspremont et al. (1983) in the context of cartel formation. An environmental agreement is said to be stable if (1) none of its members wants to leave, and (2) none of the nonmembers wants to join the agreement.

Definition 3 (Stable coalition) *Coalition $S \subseteq N$ is said to be stable if it is (1)* internally stable*: $W_i^S \geq W_i^{S \setminus \{i\}}$ for all $i \in S$ and (2)* externally stable*: $W_j^S \geq W_j^{S \cup \{j\}}$ for all $j \in N \setminus S$.*

If some member of S could obtain a higher pay off by leaving the coalition, the coalition S can no longer be sustained. However, one might wonder whether the threat of a member of S to leave is credible. Assume for instance that the first insider is worse off being a member of S compared to leaving the coalition: $W_1^S \leq W_1^{S \setminus \{1\}}$ where $W_1^{S \setminus \{1\}}$ denotes lifetime utility of player 1 under the PANE w.r.t. coalition $S \setminus \{1\}$. We say in this case that country 1 has an *objection* against coalition S. Notice that we assume that when player 1 has left coalition S, the remaining players of the coalition remain together and re-optimize their strategies according to the PANE w.r.t. to the new coalition $S \setminus \{1\}$. However, is this objection by 1 credible? Assume that there is another insider, say 2 for whom: $W_2^{S \setminus \{1\}} \leq W_2^{S \setminus \{1,2\}}$. Hence, once country 1 has defected from the coalition S, country 2 has an objection against the remaining coalition $S \setminus \{1\}$ and wants to leave the coalition. And perhaps player 3 has an objection against coalition $S \setminus \{1,2\}$ and so on. This has important implications for country 1 while considering its original deviation. It would be a myopic strategy by 1 to base its decision to leave S only upon its expected pay off $W_1^{S \setminus \{1\}}$. If player 1 is farsighted, he will take into account the possibility that there might come subsequent deviations from coalition $S \setminus \{1\}$. Ideally, player 1 should base its decision to stay in or to leave coalition S on the pay off it can obtain in the final stage of the deviation chain. The following definitions, based upon the concept of the largest consistent set by Chwe (1994) capture this idea.

4.2 Farsighted coalitional stability

First we define an *inducement relation.* A coalition S can induce a subcoalition $T \subset S$ if there exists a sequence of players such that T can be formed by successive profitable deviations by these players from coalition S.

Definition 4 (Inducement) *Coalition $S \subseteq N$ can* induce *coalition $T \subset S$ if there exists a finite sequence of players $\sigma = \{i_1, i_2, \ldots, i_m\}$, all members of S, such that $T = S \setminus \{i_1, i_2, \ldots, i_m\}$ and*

$$W_{i_k}^{S\setminus\{i_1,i_2,\ldots,i_k\}} \geq W_{i_k}^{S\setminus\{i_1,i_2,\ldots,i_{k-1}\}} \qquad \forall i_k \in \sigma$$

Basically this notion of inducement says that coalition T can be reached, starting from coalition S, by successive profitable deviations by individual members of S. Using this inducement notion, we can define a credible objection of a player against a coalition.

Definition 5 (Credible objection) *Player $i \in S \subseteq N$ has a* credible objection *against coalitions S if there exists a subcoalition $T \subset S$ such that*
(1) S can induce T,
(2) $W_i^T > W_i^S$, and
(3) none of the members of T has a credible objection against T.
No player can have an objection against the trivial coalition structure.

Objection W_i^T of a player $i \in S$ is credible only if three conditions are fulfilled simultaneously. First, the move from S to T should be possible according to the inducement relation, i.e. there is some chain of profitable deviations leading from S towards T. Secondly, player i should be better off under T than under S, and thirdly, no player of T should have a credible objection against T itself. This definition seems complicated because of its recursive nature. But this recursivity defines some kind of consistency requirement for deviations. Since no player can have an objection against the trivial coalition structure, we can always resolve the recursivity in the definition. Finally, we can define a farsighted stability for coalitions.

Definition 6 (Farsighted stability) *Coalition $S \subseteq N$ is stable in the farsighted sense if none of the member of S has a credible objection against S.*

Since we assumed that no player can object against the trivial coalition structure, it follows that this coalition structure (hence the noncooperative Nash equilibrium) is stable in the farsighted sense. The interesting question is of course whether there exist larger coalition structures that are stable in the farsighted sense. In the following section we illustrate the concept of farsighted stability for the Kyoto Protocol on greenhouse gas emissions.

5 Simulations for the Kyoto Protocol

5.1 The Eyckmans and Tulkens (1997) simulation model

We now turn to the simulation part. All of these simulations were carried out by means of a stylized dynamic simulation model described in Eyckmans and Tulkens (1999). This model is a slight adaptation of the well established RICE model by Nordhaus and Yang (1996). The RICE model is an multiregion dynamic growth model that includes a simple representation of the global climate system. RICE divides the world into 6 major regions: $N = \{USA, Japan, EU, China, FSU, ROW\}$. In every region, production is a function of capital and labour input. Capital accumulation is endogenous and labour force growth is exogenous. Technological change is modelled in a Hicks neutral way. Carbon emissions are assumed proportional to total production and the emission-output ratio is exogenously declining over time as result of Autonomous Energy Efficiency Improvement. Accumulation of atmospheric carbon is modelled as a standard capital accumulation process. The atmospheric carbon concentration drives radiative forcing causing sea level and surface temperature changes.

There are three important differences between the model in Eyckmans and Tulkens (1999) and the original RICE model by Nordhaus and Yang (1996). First, Eyckmans and Tulkens (1999) assume that all regions are autarkic, i.e. there is no trade in consumption. Secondly, utility is assumed linear in Eyckmans and Tulkens (1999) instead of logarithmic in order to allow for a transferable utility representation of the global carbon emission game. Thirdly, damage estimates and the discount rate are revised by Eyckmans and Tulkens (1999) such that future climate change damages weigh more heavily in the current period objective function. This was done in order to make the equilibrium emission reduction for the Kyoto group in the simulation model consistent with the overall emission reduction target for Annex B countries agreed in the 1997 Kyoto Protocol. Hence, we recalibrated the parameters such that the actual Kyoto Protocol is a PANE w.r.t. the Annex B coalition in our simulation model. A full listing of the equations and parameter values of the simulation model are given in appendix. The differences w.r.t. the original formulation of RICE and justification for the changes are discussed in Eyckmans and Tulkens (1999).

5.2 No transfers among coalition members

5.2.1 Myopic internal stability analysis

Let us start with the scenario without transfers and check whether the Kyoto coalition is internally stable in the sense of definition 3. This means that we have to check whether every player is better off being part of the grand coalition compared to the pay off he would get by leaving the agreement and free riding. Table 1 contains the pay offs of the Kyoto members under the Protocol and compares these with their pay off (in billion 1990US$) from free riding while assuming that the remaining coalition stays together and reoptimizes, i.e. under the PANE w.r.t. coalition $S \setminus \{i\}$.

Table 1: Instability of Kyoto, no emission trading

region	W_i^K	$W_i^{K\setminus\{i\}}$	$W_i^{K\setminus\{i\}} - W_i^K$	%
USA	111469	111483	14	0.013
Japan	61341	61345	4	0.007
EU	147034	146968	-66	-0.045
FSU	34896	34992	96	0.275

For the Kyoto coalition members, only *EU* has a clear incentive to stay in the coalition ($W_i^{S\setminus\{i\}} < W_i^S$). All other members can improve their pay off by leaving the coalition and hoping that the remaining group will continue its efforts ($W_i^{S\setminus\{i\}} > W_i^S$). However, the differences are rather small. The *FSU* has the strongest incentive to free ride but the gain amounts only a quarter of a percent. Still, according the the stable coalition concept, the Kyoto group is internally unstable since most of the members would like to leave.

Why do some regions win from free riding? Since there are no transfers, winning or loosing has to do only with the comparison of abatement costs and climate change damages. For instance, *FSU* has very low marginal abatement costs which implies that it is asked to perform a lot of emission reduction in the cooperative Kyoto Protocol. Since *FSU* does not value much the reduction in climate change damages, it ends up worse off joining the Kyoto Protocol compared to free riding.

5.2.2 Farsighted stability analysis

Deviation by USA

We start by analyzing a possible deviation of *USA* from the Kyoto coalition. According to Table 1, *USA* could benefit from defecting from the Kyoto coalition provided the three remaining members of the Kyoto group continue to cooperate. The question is of course whether this is a realistic assumption, do *Japan*, *EU* or *FSU* have no incentives to defect from the remaining coalition? Whenever some of the remaining regions would like to leave the Kyoto coalition as well, *USA* should consider its pay off under these subsequent subdeviations to evaluate its free riding incentive.

Table 2 summarizes the relevant pay offs for subsequent subdeviations from the coalition

$K \setminus \{USA\}$. This type of table will be used frequently in the sequel of the paper and therefore it is useful to explain precisely what it contains. Every line of the table contains the pay off for the Kyoto member regions under PANE w.r.t. different coalitions. The first column contains a key that describes the composition of the coalition. A "1" means that the corresponding region is member of the coalition, a "0" means that it does not belong to that coalition[6]. Hence, key "111010" refers to coalition (*USA*, *Japan*, *EU*, *FSU*) and corresponds roughly to the Annex B countries of the Kyoto Protocol. We will denote the Kyoto coalition by capital K. The key "010100" refers to the couple (*Japan*, *China*) and so on. Finally, key "000000" refers to the trivial coalition structure where there is no cooperation and all countries act as singletons.

The first line in Table 2 shows the pay off for the different members of the Kyoto group in the PANE w.r.t. coalition "111010" without any transfers or emission trading. It is the reference case of cooperation among the four Kyoto member regions. The second line (coalition key "011010") contains the pay off for the PANE w.r.t. the three player coalition "011010", i.e. the remaining coalition after *USA* has left the original Kyoto coalition "111010". Hence, it is assumed in line 2 that the three remaining regions continue to cooperate but they reoptimize their emission strategies among themselves. According to Table 2, *USA* can improve its pay off by defection from the Kyoto coalition provided the remaining coalition stays together.

Table 2: Stability analysis for $S = \{Japan, EU, FSU\}$

coalition	USA	Japan	EU	FSU
111010	111469	61341	147034	34896
011010	111483	61308	146948	34923
001010	111452	61306	146920	34937
010010	111423	61288	146888	34961
011000	111415	61276	146862	34972
000000	111398	61277	146852	34966

But this is probably not the end of the deviation. Lines 3 through 5 look one step further in the deviation process. Is in in the interest of *Japan*, *EU* and *FSU* to stay in coalition "011010"? Line 3 considers a deviation by *Japan* (coalition key "001010"), Line 4 considers a deviation by *EU* (coalition key "010010"), and Line 5 considers a deviation by *FSU* (coalition key "011000"). Still this might not be the end of the deviations chain. The remaining two player coalitions in lines 3 through 5 might disintegrate even further by deviations by one of its members. In the case of such a deviation, we end up in the last line representing pay offs of the regions in the trivial coalition structure "000000", i.e. the noncooperative Nash

[6] Recall that we have six regions ordered in the following way: $N = (USA,\ Japan,\ EU,\ China,\ FSU,\ ROW)$.

equilibrium where there is no cooperation at all.

In order the evaluate the profitability of a deviation by *USA* from the 4 player Kyoto coalition "111010", we have to check all possible patters of subsequent deviations from the 3 player coalition "011010". Let us start with a deviation by *Japan*. It turns out that *Japan* cannot gain by deviating from the three player coalition "011010", no matter what subsequent deviations might still occur. Indeed, even if either *EU* or *FSU* would deviate from the subsequent coalition (and hence ending up in the trivial coalition structure "000000"), *Japan* would always loose compared to coalition "011010". Likewise for *EU*. No matter what subsequent subdeviations, *EU* cannot by loose from leaving the 3 player coalition "011010".

Things are more complicated for *FSU* however. If *FSU* leaves the 3 player coalition "011010", it can achieve 34972 as a pay off (see coalition "011000") which is strictly better than 34923. But the 2 player coalition "011000" is not internally stable either since *Japan* can achieve a higher pay off in the trivial coalition structure "000000". This implies that when *FSU* considers deviating from "011010", it should look not at its direct pay off under "011000", but at its pay off in the final stage of the deviation chain. *FSU* can be sure that its deviation will trigger of a further (and final) subsequent deviation by *Japan*. Hence, the expected pay off for *FSU* from deviating is given by its noncooperative Nash pay off (coalition structure "000000") which amounts to 34966. Still, this is strictly higher then its pay off under coalition structure "011010". We can conclude that, even if *FSU* is farsighted and if the game has reached the stage of coalition "011010", *FSU* will indeed deviate!

What does this imply for the original deviation by *USA*? When *USA* leaves the Kyoto coalition "111010", it can be sure that neither *Japan*, nor *EU* will deviate further from the 3 player coalition "011010". However, the deviation threat by *FSU* is credible since this region can always gain from free riding, no matter what subsequent subdeviations are still to come. Therefore, if *USA* is farsighted, it will compare its pay off in the Kyoto coalition (111469) with its noncooperative Nash pay off (111398). This comparison shows that *USA* is bound to lose from leaving the Kyoto Protocol, the deviation threat by *USA* should be labeled as not credible.

Deviation by Japan

We have to repeat this analysis for the following Kyoto Protocol member, i.e. *Japan*. The deviations by *USA* and *EU* from the three player coalition "101010" are not credible since they would both lose no matter what subsequent deviations are still to come. Again the deviation by *FSU* is credible. Its deviation will cause *USA* to deviate further so that we end up in the trivial coalition structure. *FSU*'s Nash equilibrium pay off (34966) is greater than what it can obtain in the 3 player coalition "101010" (34913). This implies that for evaluating

Table 3: Stability analysis for $S = \{USA, EU, FSU\}$

coalition	USA	Japan	EU	FSU
111010	111469	61341	147034	34896
101010	111456	61345	146996	34913
001010	111452	61306	146920	34937
100010	111434	61301	146918	34952
101000	111391	61303	146892	34982
000000	111398	61277	146852	34966

the expected pay off of its possible deviation from the Kyoto Protocol, *Japan* should compare its welfare under the Protocol (61341) with what it gets in the noncooperative Nash solution (61277). We conclude that the original deviation by *Japan* from the Kyoto coalition "111010" should be labelled as noncredible. Subsequent deviations will lead to the total disintegration of the original Kyoto coalition.

Deviation by EU

Table 4: Stability analysis for $S = \{USA, Japan, FSU\}$

coalition	USA	Japan	EU	FSU
111010	111469	61341	147034	34896
110010	111448	61310	146968	34942
010010	111423	61288	146888	34961
100010	111434	61301	146918	34952
110000	111401	61281	146881	34973
000000	111398	61277	146852	34966

We turn to *EU*. *EU*'s pay off under the Kyoto Protocol amounts to 147034US$. If it were to defect from the Protocol, it cannot but lose, independent of the subsequent deviations. Clearly, it would be irrational for *EU* to deviate.

Deviation by FSU

Finally, we turn to *FSU*. The analysis for *FSU* shows that whatever subsequent deviations might occur, *FSU* will always win from leaving the Kyoto Protocol. Indeed, its pay off under Kyoto amounts to 34896 and the worst possible outcome under subsequent deviations by the

Table 5: Stability analysis for $S = \{USA, Japan, EU\}$

coalition	USA	Japan	EU	FSU
111010	111469	61341	147034	34896
111000	111394	61297	146916	34992
011000	111415	61276	146862	34972
101000	111391	61303	146892	34982
110000	111401	61281	146881	34973
000000	111398	61277	146852	34966

remaining players is 34966, *FSU*'s pay off under the noncooperative Nash equilibrium. No matter who deviates after *FSU* has left Kyoto and no matter where the deviation process ends up, it is always in the interest of *FSU* to defect.

Summarizing the farsighted stability analysis without transfers, we see that only one region, *FSU*, has a credible threat to opt out of the Kyoto Protocol. The other regions cannot credibly commit to deviating since subsequent deviations of remaining members will lead to a worse outcome. The final appreciation is that the Kyoto Protocol is not internally stable in the farsighted sense, the problematic region is *FSU*. It is interesting to compare this result with the outcome of the myopic stability analysis. The Kyoto Protocol is internally instable in the myopic sense and three regions out of four were identified as problematic w.r.t. coalitional stability. The contribution of the farsighted stability analysis is that it reduces the number of deviation threats, since deviations have to be credible, and that it increases the probability of finding a stable coalition.

5.2.3 Is there perhaps a smaller coalition that is stable in the farsighted sense?

The analysis above indicates that *FSU* has strong incentives to deviate from the Kyoto coalition. But what about the remaining coalition "111000"? Does the fact that *FSU* wants to deviate say something about the farsighted stability of the remaining coalition? This is hard to say beforehand, the only way to answer this question is to repeat the analysis for coalition "111000". This can be done easily by inspection of Table 5. *USA* has a credible objection against coalition "111000" because it can be sure of its noncooperative pay off (111398) which is better than what it gets in the PANE w.r.t. coalition "111000" (111394). Hence, the three player coalition (*USA*, *Japan*, *EU*) is not stable in the farsighted sense. This line of reasoning can be continued to show that when both *FSU* and *USA* have left the Kyoto Protocol, the remaining couple is not stable in the farsighted sense either. In particular, *Japan* has a credible objection against coalition "011000". In the end, complete disintegration of the Kyoto Protocol seems unavoidable.

5.3 Emission trading among the Kyoto coalition members

5.3.1 Myopic internal stability analysis

One might argue that the negative result on the stability of the Kyoto Protocol was to be expected since no transfers are used to compensate high effort countries within a coalition. *FSU* is characterized by the lowest marginal abatement costs in the Kyoto coalition and is therefore required to perform a lot of abatement effort under the PANE w.r.t. the Kyoto coalition. Without compensation, *FSU* is worse off being a member of the coalition compared to its free riding strategy. Moreover, the actual Kyoto Protocol includes provisions for some system of emission trading among the Annex B countries. Hence, we need to repeat the analysis while allowing for emission trading. Results are summarized in Table 6. The figures for T_i are total discounted lifetime net transfers resulting from emission trading. They correspond to the transfers in Figure 2 in Appendix.

Table 6: Instability of Kyoto, emission trading

region	W_i^K	T_i	$W_i^K + T_i$	$W_i^{K\backslash\{i\}}$	$W_i^{K\backslash\{i\}} - W_i^K$
USA	111469	92	111561	111483	-78
Japan	61341	-66	61275	61345	70
EU	147034	-137	146897	146968	71
FSU	34896	111	35007	34992	-15

The effect of allowing for trade based upon grandfathering permits is that *USA* and *FSU* are net sellers of emissions whereas all other regions are net buyers. Hence, transfers are flowing from *Japan* and *EU* towards *USA* and *FSU*. This makes the Kyoto Protocol more attractive to the latter countries. After trading, *USA* and *FSU* are better off with the Kyoto Protocol compared to their free riding strategy. Of course, the trading makes the grand coalition less attractive for the net buyers of permits. The free riding incentive becomes stronger for *Japan* and *EU* becomes a net losers from the Protocol. Again we conclude that the Kyoto coalition is not internally stable in a myopic sense. However, the pattern of winners and losers is completely different from the no transfer case.

5.3.2 Farsighted internal stability analysis

We summarize in Table 7 all the pay off figures for the different Kyoto members under the PANE w.r.t. all possible subcoalitions. All these figures take into account an emission trading scheme with grandfathering of permits and with an initial assignment corresponding to the original 1997 Kyoto Protocol.

Table 7: Stability analysis with emission trading

coalition	USA	Japan	EU	FSU
111010	111461	61275	146897	35007
011010	111483	61271	146879	35028
101010	111516	61345	146867	34983
110010	111470	61265	146968	34965
111000	111532	61249	146826	34992
110000	111435	61248	146881	34973
101000	111481	61303	146801	34982
100010	111435	61302	146918	34951
011000	111415	61267	146871	34972
010010	111423	61271	146888	34979
001010	111452	61306	146856	35001
000000	111398	61277	146852	34966

Again, the farsighted stability analysis yields a negative result. There is one region with a credible deviation, namely *Japan*. The reason being that under the trading scheme, *Japan* is an important net buyer of permits (since its domestic marginal abatement cost is high). Therefore, the Kyoto Protocol with emission trading requires *Japan* to pay for a large part of the transfers that accrue to *USA* and *FSU*. The other regions have no credible objection against the Kyoto Protocol. In particular *FSU* receives a substantial compensation for the emission abatement it produces under the Protocol since it is the major supplier of emission permits. It never wants to opt out of the Protocol and forego the important emission trading benefits.

In this case, it is worthwhile considering the three player coalition "101010" that remains after *Japan* would have left. This coalition turns out to be stable in the farsighted sense! Indeed, neither *USA* nor *FSU* have a credible objection since they are bound to lose no matter what subsequent deviation might follow. This is intuitive since they receive some transfer from *EU* under the emission trading scheme. But also *EU* lacks a credible deviation strategy. It seems to have a profitable deviation strategy in the myopic sense (it obtains 146867 under coalition "101010" against 146918 under coalition "100010") but its deviation would lead *FSU* to deviate further. Hence, when *EU* deviates we end up in the noncooperative Nash equilibrium which is less interesting for *EU* compared to its pay off under "101010".

We conclude that the stability analysis is affected in a substantial way by allowing for emission trading. The pattern of winners and losers is different but the full Kyoto coalition remains unstable since *Japan* has a credible deviation strategy. However, once *Japan* has left the coalition, the remaining parties *USA*, *EU* and *FSU* have an interest in staying together. When they reoptimize (according to the PANE w.r.t. "101010") and allow for emission trading (with grandfathering permits according to the original Kyoto Protocol emission assignments), they can form a stable coalition in the farsighted sense.

6 Conclusion

In this paper we explore the coalitional stability of the 1997 Kyoto Protocol on the emissions of greenhouse gases. Simulations with a stylized integrated assessment model suggest the following conclusions. (1) The surplus sharing rule within a coalition is very important to the final coalition structure that can form. In particular, allowing for emission trading and grand fathering initial permits seems to provide better opportunities for stability than no transfers at all. (2) Conventional myopic stability analysis suggests that several signatories of the Kyoto Protocol would have a profitable free riding strategy. (3) Introducing farsightedness strongly restricts the number of credible free riding strategies. In particular, only the Former Soviet Union seems to have a potentially credible deviation strategy in the absence of transfers. If emission trading is allowed for, only Japan has a credible deviation. (4) With emission trading based on the Kyoto Protocol emission assignments, the coalition defined by the Annex B countries minus Japan is stable in the farsighted sense.

This analysis is only a first step in the appraisal of the stability of the Kyoto Protocol. First, the stylized simulation model can be improved both on the economics and on the physical part. Given the often small differences in pay off for coalition members, the uncertainty on emission reduction cost and climate change damage parameters seriously affects the conclusions. Secondly, the concept of farsighted stability needs further elaboration in order to assess its strengths and flaws. In particular, it would be very useful to apply the concept in a theoretical model with identical countries and to compare it to competing coalitional stability concepts like, for instance, in Finus and Rundshagen (1998). Thirdly, instead of focussing on the Kyoto Protocol, it would be interesting to look at all other possible coalition structures and to identify the largest consistent set, as defined by Chwe (1994), in the global warming game.

References

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

Barrett, S. (1997), Towards a theory of international cooperation, in: Carraro, C. and Siniscalco, D. (eds), *New directions in the economic theory of the environment* (Cambridge University Press, Cambridge)

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

Carraro, C. (1999), The structure of international environmental agreements, in: Carraro, C. (ed.) *International environmental agreements on climate change* (Kluwer Academic Publishers, Dordrecht), 9–25

Chander, P. and Tulkens, H. (1995), A core-theoretic solution for the design of cooperative agreements on transfrontier pollution, *International Tax and Public Finance* **2**, 279–293

Chander, P. and Tulkens, H. (1997), The core of an economy with multilateral environmental externalities, *International Journal of Game Theory* **26**, 379–401

Chwe, M. S.-Y. (1994), Farsighted coalitional stability, *Journal of Economic Theory* **63**, 299–325

d'Aspremont, C., Jacquemin, A., Gabszewicz, J.J., and Weymark, J. (1983), On the stability of collusive price leadership, *Canadian Journal of Economics* **16**, 17–25

Eyckmans, J., and Tulkens, H. (1999), Simulating with RICE coalitionally stable burden sharing agreements for the climate change problem, CORE Discusion Paper 9926

Finus, M., and Rundshagen, B. (1998), Towards a positive theory of coalition formation and endogenous instrumental choice in global pollution control, *Public Choice* **96**, 145–186

Nordhaus, W.D. and Yang, Z. (1996), A regional dynamic general-equilibrium model of alternative climate-change strategies, *American Economic Review* **86**, 741–765

Samuelson, P.A. (1954), The pure theory of public expenditures, *Review of Economics and Statistics* **36**, 387–389

Appendix

A simplified version of the RICE model

For the purpose of the simulation exercise, we used a simplified version of the RICE model, originally developed by Nordhaus and Yang (1996). A complete list of the equations of the simplified model is given below:

$$W_i(Z_{i,t}) = \sum_{t=0}^{T} \frac{Z_{i,t}}{[1+\rho_i]^t} \tag{29}$$

$$Y_{i,t} = Z_{i,t} + I_{i,t} + C_{i,t} + D_{i,t} \tag{30}$$

$$Y_{i,t} = A_{i,t}\, K_{i,t}^{\gamma}\, L_{i,t}^{1-\gamma} \tag{31}$$

$$C_{i,t} = Y_{i,t}\, a_{i,1}\, \mu_{i,t}^{a_{i,2}} \tag{32}$$

$$D_{i,t} = Y_{i,t}\, b_{i,1}\, \Delta T_t^{b_{i,2}} \tag{33}$$

$$K_{i,t+1} = [1 - \delta_K]\, K_{i,t} + I_{i,t} \qquad K_{i,0} \text{ given} \tag{34}$$

$$E_{i,t} = \sigma_{i,t}\, [1 - \mu_{i,t}]\, Y_{i,t} \tag{35}$$

$$M_{t+1} = [1 - \delta_M]\, M_t + \beta \sum_{i \in N} E_{i,t} \qquad M_0 \text{ given} \tag{36}$$

$$F_t = \frac{4.1\ \ln(M_t/M_0)}{\ln(2)} + F_t^x \tag{37}$$

$$T_t^o = T_{t-1}^o + \tau_3\, [T_{t-1}^a - T_{t-1}^o] \tag{38}$$

$$T_t^a = T_{t-1}^a + \tau_1[F_t - \lambda T_{t-1}^a] - \tau_2[T_{t-1}^a - T_{t-1}^o] \tag{39}$$

Calibrating RICE to the 1997 Kyoto Protocol emission targets

The Kyoto group or Annex B countries

Consider the coalition $K = \{\mathit{USA}, \mathit{Japan}, \mathit{EU}, \mathit{FSU}\}$ which approximates most closely the group of Annex B countries of the 1997 Kyoto Protocol. The reader familiar with the composition of list of Annex B countries to the Kyoto Protocol will see that coalition K is only a crude approximation. In particular, some OECD countries like Australia and New Zealand did sign the Kyoto Protocol but are member of the ROW region in RICE. In addition, some of the former Eastern European countries that signed the Kyoto Protocol are not included in

Table 8: List of variables

$Y_{i,t}$	production
$A_{i,t}$	productivity
$Z_{i,t}$	consumption
$z_{i,t}$	per capita consumption
$I_{i,t}$	investment
$K_{i,t}$	capital stock
$L_{i,t}$	population
$C_{i,t}$	cost of abatement
$D_{i,t}$	damage from climate change
$E_{i,t}$	carbon emissions
$\sigma_{i,t}$	emission-output rate
$\mu_{i,t}$	emission abatement
M_t	atmospheric carbon concentration
F_t	radiative forcing
T_t^a	temperature increase atmosphere
T_t^o	temperature increase deep ocean

coalition K. They are in ROW. Still, overall, the composition of the group is consistent with the list of Annex B countries.

Additional model modifications
Several additional modifications had to be made to the model by Eyckmans and Tulkens (1999) in order to be able to simulate the Kyoto Protocol as a PANE w.r.t. coalition K. The first modification concerns the question what will happen after 2012, i.e. after the compliance period of the Kyoto Protocol. The Protocol only stipulates emission reduction targets to be reached in the compliance period $2008 - 2012$. We want to simulate on a much longer time horizon. Hence we have to make some assumption on what comes after Kyoto. We would argue that the best we can do is to assume that the Kyoto group determines its emission abatement path as to maximize its joint lifetime consumption. This is basically saying that we will use the PANE concept for calculating the optimal strategy for the Kyoto group. However, if we run the simulation model with the original parameter values used by Nordhaus and Yang (1996), it turns out that the 5% emission reduction by $2008 - 2012$ agreed upon by the Annex B countries is more than what would be optimal according to their PANE w.r.t. K. In order to make the PANE prediction compatible with the real world emission targets of the Kyoto Protocol, I therefore revised several damage related parameters of the model. In particular, the damage estimates for the Kyoto group are revised upward and their discount rates (1% instead of 3%) are lowered.

Secondly, the protocol accommodates for several flexible mechanisms like bubbles and emission trading with, possibly, caps. We use the emission trading scenario with grandfathering as described higher. In reality, the flexible mechanisms will probably not attain complete cost efficiency. In that sense, the results reported here are overly optimistic. For the initial assignment of carbon permits, we used the distribution based on the emission objectives in

Table 9: List of parameters

ϵ	inequality aversion	0
δ_K	capital depreciation rate	0.10
γ	capital productivity parameter	0.25
β	airborne fraction of carbon emissions	0.64
δ_M	atmospheric carbon removal rate	0.0833
τ_1	parameter temperature relationship	0.226
τ_2	parameter temperature relationship	0.44
τ_3	parameter temperature relationship	0.02
λ	parameter temperature relationship	1.41
M_0	initial carbon concentration	590
T_0^a	initial temperature atmosphere	0.50
T_0^o	initial temperature deep ocean	0.10

Table 10: Parameter values

	$b_{i,1}$	$b_{i,2}$	$a_{i,1}$	$a_{i,2}$	ρ_i	$\mu_{i,2010}$
USA	0.01102	3.0	0.07	2.887	0.01	0.07
Japan	0.01174	3.0	0.05	2.887	0.01	0.06
EU	0.01174	3.0	0.05	2.887	0.01	0.08
China	0	3.0	0.15	2.887	0.03	-
FSU	0.00857	3.0	0.15	2.887	0.01	0.00
ROW	0	3.0	0.10	2.887	0.03	-

the Kyoto Protocol. The relative share of a Kyoto member in total Kyoto emission for the compliance period 2008-2012 is used to distribute emission titles in all subsequent periods. We are aware of the restrictive nature of this assumption since one might argue that the permit allocation rule will probably change over time. Since there is no information on post-Kyoto emission targets we had to make some simplifying assumption.

Thirdly, the regions *China* and *ROW* are real heavy-weights, both in terms of populations as in terms of emissions. *ROW* consists of all developing countries (among others Brazil, India, Indonesia), all oil producing countries in the Gulf area, South Africa, Israel, Australia and New Zealand (nonexhaustive list). For both regions we made the simplifying assumption that they do not value climate change damages. For *China*, this assumption is based more on observation of the climate change negotiations in Kyoto. Basically, *China* and most developing countries refused to join the Protocol because they argued that industrialized countries should take the lead in abating carbon emissions since they are responsible for the bulk of past emissions. Moreover they refused to sign a Protocol which might hinder their future growth prospects. In order to approximate developing countries' behaviour, we therefore assume that *China* and *ROW* do not value climate change damages. For *ROW* this can be justified by

Table 11: Parameter values

	Y_i^0	K_i^0	L_i^0	E_i^0
USA	5464.796	14262.51	250.372	13.60
Japan	2932.055	8442.25	123.537	2.92
EU	6828.042	18435.71	366.497	8.72
China	370.024	1025.79	1133.683	6.69
FSU	855.207	2281.90	289.324	10.66
ROW	4628.621	9842.22	3102.689	17.00

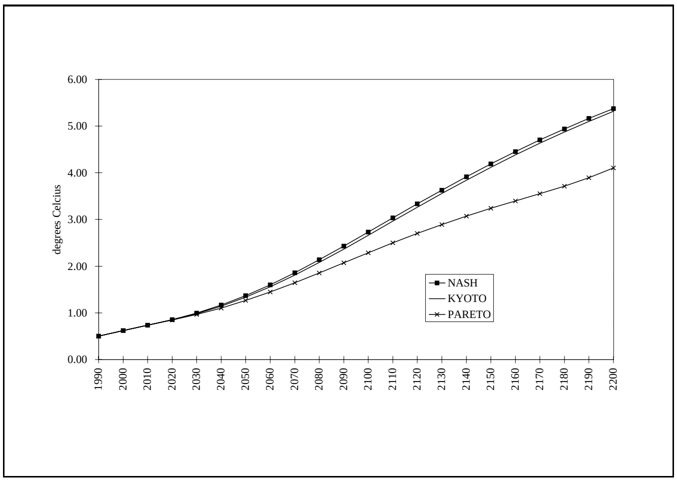
still another argument. Since this region is highly diverse it would be unrealistic to assume that its member states perfectly internalize all damages among themselves. It turns out that this assumption does not change the overall picture of the simulations.

Fourthly, compared to the Nordhaus and Yang (1996) paper we revised the exogenous productivity growth rate for *FSU* in the first periods of the planning horizon downward. As a result Russian production is lower than in Nordhaus and Yang (1996), in fact they experience a decline in production in the first periods. This result is more in line with the observed output fall during the 1990ies in *FSU*. In this respect it is also important to mention that under our revision, baseline carbon emissions in 2010 of *FSU* without any emission abatement coincide roughly with their 1990 emission level. This means that *FSU* can meet its Kyoto Protocol target without any abatement measures and that it has no so-called “hot air” to sell. All permits sold by *FSU* correspond to actual emission abatement activities.

Reference run simulations
Let us briefly look at the outcome of the PANE w.r.t. K in terms of two key variable, nl. global mean temperature increase and the transfers implied by the emission trading with grandfathering. Figure 1 shows the evolution of global mean surface temperature under three different scenario's: complete absence of cooperation (NASH), the PANE w.r.t. the Kyoto coalition (KYOTO) and complete cooperation (PARETO). Notice that there is very little impact of the Kyoto agreement, even if it is continued after 2012 in the form of a PANE, on global mean temperature. The difference with the noncooperative scenario is hard to distinguish. The difference with the cooperative scenario is substantial. The intuition for this lack of impact is clear if one looks at the evolution of the share of the Kyoto group in total carbon emissions. Their share declines from more than 60% in 1990 to about 16% in 2100. A huge increase in emissions by *China* and *ROW* is to be expected and this rise is unchecked by the Kyoto Protocol provisions.

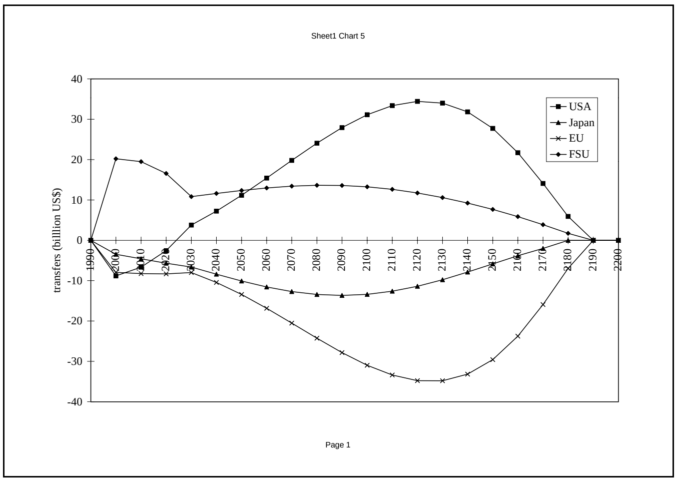
Figure 2 shows the time path of net transfers resulting from emission trading among the Kyoto member regions. Recall that for all periods, total emission reduction for the Kyoto group corresponds to its PANE. The distribution of the permits is made in proportion to the assignment of emissions for the compliance period 2008-2012 in the Kyoto Protocol. The corresponding weights are: 0.37 for *USA*, 0.08 for *Japan*, 0.24 for *EU* and 0.31 for *FSU*. The figure shows that over the entire horizon, *Japan* and *EU* are net buyers of permits and that *FSU* is a net seller. Interestingly, *USA* is a net buyer initially but it becomes a net seller

Figure 1: Atmospheric temperature



of permits from 2030 onwards. The price of the permits equals the marginal abatement cost (which is equal in all Kyoto member regions). It increases steadily from about 60US$ in 2010 to approximately 110US$ by the year 2100. Average abatement w.r.t. BAU emissions amounts to 23% for the Kyoto coalition.

Figure 2: Emission trading transfers



The Center for Economic Studies (CES) is the research division of the Department of Economics of the Katholieke Universiteit Leuven. The CES research department employs some 100 people. The division Energy, Transport & Environment (ETE) currently consists of about 15 full time researchers. The general aim of ETE is to apply state of the art economic theory to current policy issues at the Flemish, Belgian and European level. An important asset of ETE is its extensive portfolio of numerical partial and general equilibrium models for the assessment of transport, energy and environmental policies.

WORKING PAPER SERIES

n° 2001-03 Eyckmans J. (2001), On the farsighted stability of the Kyoto Protocol

n° 2001-02 Van Dender, K. (2001), Pricing transport networks with fixed residential location

n° 2001-01 Rousseau, S. and Proost, S. (2001), The relative efficiency of environmental policy instruments in a second-best setting with costly monitoring and enforcement (*also available as CES Discussion Paper 01.04*)

n° 2000-09 Proost, S., and Van Regemorter, D. (2000), How to achieve the Kyoto Target in Belgium — modelling methodology and some results

n° 2000-08 Eyckmans J. and Bertrand, C. (2000), Integrated assessment of carbon and sulphur emissions, simulations with the CLIMNEG model

n° 2000-07 Pepermans G., and Proost, S. (2000), Stranded costs in the electricity sector

n° 2000-06 Calthrop, E., and Proost, S. (2000), Regulating urban parking space: the choice between meter fees and time restrictions (*also available as CES Discussion Paper 00.21*)

n° 2000-05 Mayeres, I., and Proost, S. (2000), Should diesel cars in Europe be discouraged? (*also available as CES Discussion Paper 00.18*)

n° 2000-04 Willems, B. (2000), Cournot Competition in the Electricity Market with Transmission Constraints (*also available as CES Discussion Paper 00.24*)

n° 2000-03 Pepermans, G., and Proost, S. (2000), The Liberalisation of the Energy Sector in the European Union

n° 2000-02 Eyckmans, J., and Cornillie, J. (2000), Efficiency and Equity of the EU Burden Sharing Agreement

n° 2000-01 Bigano, A. (2000), Environmental Regulation for the Liberalised European Electricity Sector. Towards a Numerical Approach