The effect of climate thresholds on coalition formation: an application of numerical models[1]

Johannes Emmerling[2], Ulrike Kornek[3], Valentina Bosetti[4], Kai Lessmann[3], Massimo Tavoni[1]

Abstract: The presence of thresholds in the climate system may crucially influence the nature of the climate game and the prospects for international cooperation. Analytical calculations show that large coalitions which keep temperatures below a sharp threshold can be stable when crossing the damage implies large damage costs and requires sufficiently low abatement costs. By applying two numerical models of coalition formation, we numerically characterize both the location of the threshold and the associated damage costs that would lead to such self-enforcing agreements. We show that because regions differ concerning their vulnerability to climate change and especially their abatement potential, these agreements do not include all regions: those which are pivotal to keeping the threshold have the largest incentive to cooperate. Confirming previous analytical literature, we show that introducing uncertainty about the location of the threshold impedes cooperation compared to the absence of uncertainty. The numerical exercise, however, allows us to show that still the presence of an uncertain threshold may improve the prospects for cooperation compared to the absence of a threshold.

1. Introduction

The 20th century has seen the rise of many international transboundary pollution problems. While international agreements have led to significant improvements of environmental quality in many areas, negotiations on global climate change mitigation have so far not achieved sufficient abatement of greenhouse gas emissions (Rogelj et all 2010). The global public good nature of abatement impedes comprehensive cooperation because free-riding on the abatement of other countries is possible while individual abatement costs are avoided (Barrett 2003). Hence, the design of an international agreement should facilitate the cooperation of countries which act non-cooperatively.

[1] Very preliminary draft. Please do not cite or circulate. Comments are welcome.
[2] Fondazione Eni Enrico Mattei (FEEM) and Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Italy
[3] Potsdam Institute for Climate Impact Research (PIK), PO Box 60 12 03, D-14412 Potsdam, Germany
[4] Department of Economis Università Bocconi and Fondazione Eni Enrico Mattei (FEEM) Milan, Italy

Early coalition formation literature has assessed that only small and ineffective coalitions are stable when countries base their decision to sign an international climate agreement on marginal changes in costs of abatement and damage costs from emissions (Hoel 1992, Barrett 1994). The recent literature has highlighted how additional means may improve the participation and environmental effectiveness of an agreement (Finus 2008, Lessmann et al. 2009). One of the most important features favoring climate cooperation is that countries differ with respect to their costs of abatement and the associated damage costs. Incorporating well-designed transfers between regions allows larger coalitions to become stable (Weikard 2009, Lessmann et al. 2014) and improves environmental quality.

Other properties of climate change beyond its public good character should be explored with respect to their implications for cooperation. While the majority of the climate coalition formation literature considers continuous damages from greenhouse gas emissions, recent studies have emphasized the role of thresholds in the climate system. Barrett (2013) shows that it can be in the self-interest of countries to keep temperatures below a climate threshold if the damage costs associated with crossing the threshold are sufficiently large and the costs of abatement are sufficiently low. Because individual countries are not able to keep the threshold by themselves and the individual costs of crossing it are too high, an international agreement is a means for countries to coordinate on the social optimum, which represents a non-cooperative Nash-equilibrium of the game in abatement strategies. The agreement does not need to provide incentives for countries to cooperate but serves as a means to coordinate if a threshold of sufficient characteristics is present.

Barrett and Dannenberg (2012) and Barrett (2013) highlight that uncertainty about the location of the threshold may again reverse the implications that the presence of thresholds has on cooperation vs. coordination. If the exact amount of abatement to avoid crossing of the climate threshold is unknown, the point of reference for coordination vanishes with increasing uncertainty. There exists a range of parameter values for which the problem of climate change may still be a coordination game but the values seem rather unrealistic.

The characteristics of climate thresholds with respect to their location and the related uncertainty, as well as the associated damage costs are hence crucial for their prospects to coordinate on the necessary abatement. Lenton et al. (2008) name several tipping elements in the earth's response to increasing concentration levels of greenhouse gases. They report several threshold temperatures to lie in the range of possible temperature rises in the 21st century – the arctic summer ice at 0.5 to 2°C or the Greenland ice sheet at 1 to 2°C temperature increase above 1980-1999 levels. Uncertainties remain large with respect to the exact location of the threshold and the actual impact that the crossing of the threshold has on economic and social systems.

In this paper we analyze the influence of climate thresholds in two numerical climate coalition formation models. The models MICA (Lessmann et al. 2009) and WITCH (Bosetti et al. 2006) are long-term optimal growth models, in which for each world region abatement costs are valued against damage costs from emissions. Our approach allows us to extend the analytical literature in several respects. First, the models incorporate non-linear utility functions and non-linear reaction-functions. Second, each world region's abatement costs and damages are empirically calibrated, allowing for realistic differences between world regions. Third, the characteristics of the threshold can be studied numerically based on the empirical foundation of the climate system and losses of GDP. Additionally, we extend the analytical literature from a static threshold to a dynamic setting, in which abatement efforts are near-term and damage costs occur in later time periods. We test the robustness of our results both by exploring different characteristics of thresholds and by comparing the two models.

We first introduce certain thresholds in each model and explore to what extend different locations and damage costs of the threshold influence optimal abatement strategies in the social optimum. The grand coalition comprising all world regions may adjust abatement differently over time. We find that depending on the characteristics of the threshold it is either optimal to (i) keep temperatures below the threshold for the entire time-horizon, (ii) only increase abatement moderately to postpone the eventual crossing of the threshold temperature in time, or (iii) not provide any additional abatement as a reaction to the threshold.

In a second step, we study the incentive to leave the grand coalition of all regions. We find that the sub-coalitions – cooperation between all regions but one – adjust their abatement strategies differently depending on the characteristics of the region that leaves. The smaller sub-coalition may still choose to keep temperatures below the threshold even if the leaving region increases its emissions. When the additional abatement required from the sub-coalition to compensate the deviation of a region is sufficiently small, the threshold is maintained. In such cases, incentives to deviate are very large as climate damages are almost unchanged for basically zero mitigation costs. Contrarily to the usual behavior of coalitions when damage costs are continuous, we show that the presence of thresholds may imply an increase in a coalition's ambition as a reaction once a region free-rides.

We are also able to identify which regions are pivotal in keeping temperatures below specific thresholds. Typically these are regions with great abatement potentials whose unwillingness to join makes the threshold unattainable. These regions might then be deterred from leaving the coalitions, if the damages of crossing the thresholds are high enough. Therefore, while the presence of thresholds increases the scope for cooperation, this asymmetry across regions calls for additional transfers that may be necessary to reach the social optimum in a stable agreement. [5]

In a last step we exemplarily simulate uncertainty with respect to the temperature at which the threshold is crossed. Confirming the literature, we find that the scope for cooperation is significantly reduced when introducing uncertainty. In expectation, the introduction of uncertainty leads to less severe changes in damages when a region leaves the grand coalition, which leads to similar free-riding behavior previously observed in the literature.

The text is structured as follows. Section 2 describes the coalition formation model we apply. Section 3 introduces a simple analytical model that clarifies the main mechanisms of the numerical models. Section 4 describes the implementation of thresholds in the numerical models, while section 5 reports results on abatement behavior and the free-riding incentive. Section 6 concludes.

2. The model of coalition formation

We study the stability of the grand coalition of all regions, denoted G, following the predominant approach of modelling the decision to join the coalition as the first stage in a one-shot cartel-formation game. Following d'Aspremont and Gabszewicz (1986), a region decides to sign the agreement in the first stage of the game, the participation stage. In the second stage, the emission stage game, regions choose economic strategies which determine the abatement of greenhouse gases. When being a signatory to the agreement, we assume that the coalition maximizes a joint social welfare function while non-signatories maximize their individual utility (this setup is similar to the Partial Agreement Nash Equilibrium (PANE) concept introduced by Chander and Tulkens 1995).[6]

[5] This however is only the case in results coming from the MICA model. The WITCH model exhibits less scope for cooperation mostly due to different representations of dynamic abatement.

[6] WITCH implements the coalitional optimum through maximization of the utilitarian sum of individual utility per region. MICA computes the coalitional optimum by solving a competitive equilibrium on international commodity markets with full internalization of the climate change externality.

Formally, the free-riding incentive can be assessed by studying the stability function φ, which is the difference in utility π of a region i when being a signatory to the agreement and being a non-signatory to the remaining coalition:

$$\varphi_i = \pi_i^m(G) - \pi_i^{nm}(G \setminus \{i\}) \tag{1}$$

If the stability function is positive, $\varphi_i > 0$, for all regions, the grand coalition G is be stable, otherwise some regions have an incentive to leave and free-ride on the coalition.

In some cases, the free-riding incentive can be positive for some regions while other regions have an incentive to sign. In this case, the regions that have an incentive to sign may compensate the other regions for their mitigation effort to stabilize the entire coalition. We apply the method described in Kornek et al. (2014) to test whether there exists a transfer mechanism between regions such that the stability function attains positive values for every region inside the grand coalition.

3. An analytical coalition formation model

Considering thresholds in the climate game changes the incentives to join an agreement crucially compared to assuming continuous damage costs from abatement (for a discussion of the underlying mechanisms in the continuous case see Karp and Simon 2013). In order to understand the basic mechanisms in more detail, this section first discusses a simple analytical framework. Section 3.1 introduces deterministic thresholds to show that depending on the parameters of the game and the reaction of non-signatories, the grand coalition of all regions may or may not be stable. As knowing the location of the threshold is a strong assumption when considering the large uncertainties surrounding tipping elements (Lenton et al. 2008), the second model introduces uncertainty in the location of the threshold. The analysis shows that – again depending on parameter values – the presence of a negotiation stage may make cooperation more likely.

3.1. Deterministic thresholds

Consider N symmetric regions interacting via a global public good. Benefits from abatement follow a step function while abatement costs are assumed to be quadratic, leading to the following utility function:

$$\pi_i = \begin{cases} 0\ -\frac{c}{2}q_i^2 & Q < \bar{Q} \\ B\ -\frac{c}{2}q_i^2 & Q \geq \bar{Q} \end{cases} \tag{2}$$

where q_i is individual abatement, $Q = \sum_i q_i$ is cumulative abatement, $\bar{Q}$ is the threshold location, B is the monetary magnitude of the threshold, and c is the slope of marginal abatement costs. The social optimum is at $q_i = \frac{Q}{N} = \frac{\bar{Q}}{N}$ and also a Nash-equilibrium in abatement strategies if $N^2B - \frac{c}{2}\bar{Q}^2 \geq 0$ (see also Barrett 2013). We will in the following assume that this inequality holds. Another Nash-equilibrium in abatement strategies is at $q_i = 0$. Hence, the game is a coordination game in the presence of a threshold.

Consider the formation of a coalition of size $N-1$. There exists multiple equilibria in abatement strategies in the second stage of the game (remember that the social optimum is still a Nash-equilibrium in emission strategies). Since only the coalition has the chance to coordinate strategies in the second stage of the game, we assume that non-signatories simply adopt a zero-abatement strategy when free-riding. Then, the coalition will keep the threshold (meaning $q_i^m = \frac{Q}{N-1} = \frac{\bar{Q}}{N-1}$) as long as

$$(N-1)^2B - \frac{c}{2}\bar{Q}^2 \geq 0. \tag{3}$$

Hence, if a region leaves the grand coalition of all regions and the sub-coalition still provides the abatement to keep the threshold, this region has a strong incentive to leave. Benefits from abatement remain at B while abatement costs drop to zero: the value of the stability function is hence negative. In conclusion, the social optimum might not be reached by a stable agreement whenever the properties of the threshold are such that it is optimal for smaller coalitions to keep temperatures below the threshold.

3.2. Uncertain thresholds

The assumption of a known location of the threshold must be rejected as unrealistic with respect to climate change. However, the introduction of uncertainty about the location of the threshold undermines cooperation in the simple abatement game of (2). Barrett (2013) shows for a slightly different utility function that the social optimum may not be a Nash-equilibrium anymore when uncertainty is introduced. We want to study if this result still holds under the presence of a negotiation stage. In particular, we are interested whether a stable coalition exists that keeps the threshold while the social optimum is not a Nash-equilibrium anymore. For this, we introduce uniform uncertainty about the location of the threshold within the interval: $[\bar{Q}-\frac{\Delta Q}{2},\bar{Q}+\frac{\Delta Q}{2}]$ to equation (2). The expected utility is:

$$E[\pi_i]=\begin{cases} 0-\frac{c}{2}q_i^2 & Q<\bar{Q}-\frac{\Delta Q}{2} \\ \frac{B}{\Delta Q}\left[Q-\bar{Q}+\frac{\Delta Q}{2}\right]-\frac{c}{2}q_i^2 & \bar{Q}-\frac{\Delta Q}{2}\leq Q<\bar{Q}+\frac{\Delta Q}{2} \\ B-\frac{c}{2}q_i^2 & Q\geq\bar{Q}+\frac{\Delta Q}{2} \end{cases}. \tag{4}$$

If we again assume that non-signatories do not contribute to abatement when the coalition keeps the threshold, the appendix shows that there are parameter values for which a stable coalition exists that keeps the threshold even though the social optimum is not a Nash-equilibrium anymore. In the social optimum under uncertainty, the unilateral decision to marginally decrease abatement induces marginal changes to the expected utility. Under coalition formation, a free-riding region induces non-marginal changes to the abatement decisions of the coalition due to the binary decision of the participation stage. The effect of an additional stage on the success of an international environmental agreement is therefore undecided: cooperation maybe enhanced but this crucially depends on parameter settings.

4. Implementation of thresholds in numerical coalition formation models

The previous section depicts by means of a simple toy model how the coalition formation process critically depends on the parameter values of the location and damage costs of a threshold. We therefore apply two empirically calibrated integrated assessment models (IAMs) to study the characteristics of thresholds: MICA and WITCH (http://www.witchmodel.org/). Both models derive economic strategies with respect to climate change mitigation from an optimal growth framework. The models combine the two level game described above with an integrated climate economy model in the second stage. Regions are heterogeneous with respect to their damage costs and costs of abatement as opposed to the symmetric case of the analytical model (for a characterization of real-world heterogeneity in the climate game see Lessmann et al 2014). A more detailed description of the numerical calibration of each model can be found in the appendix.

The Model of International Climate Agreements (MICA, Lessmann et al., 2009) follows the same economic framework as RICE (Nordhaus and Yang 1996) but with different assumptions about abatement costs and damage costs. It relies on stylized abatement cost functions to model emissions reductions and neglects inertias in investing in abatement technologies. In contrast, WITCH incorporates an explicit representation of mitigation options, particularly in the energy system (Bosetti et al., 2006).

Thresholds enter the models through their usual implementation of damage costs. The loop between the environment and the economy is closed by a Nordhaus-type damage function that translates temperature increase to percentage losses of GDP (Nordhaus 1994). For the implementation of thresholds in the climate system, the original damage function was kept:

$D(i,t) = \Omega(i, T(t)) * GDP(i,t),$

with $\Omega(i, T(t))$ damages as a share of GDP for region i depending on the atmospheric temperature at time t and $GDP(i,t)$ production in monetary units. In the base specification, the function is continuous and moderately slopes upward in temperature for both models. Damages are deducted from production in the budget equation, which is standard in the literature.

4.1. Deterministic Thresholds

The following additional threshold-like function was added to $\Omega(i,t)$, inducing damage costs to be in accordance with the simple utilities of equation (2), however with additional continuous damage costs. Thresholds were approximated via the cumulative distribution function of the normal distribution (the error-function) due to computational reasons. The following damages were added to the original $D(i,t)$:

$d * errorfunction\left(\frac{T(t)-T_s}{\sigma}\right) * GDP(i,t),$

where: d is the maximum damage in share of GDP from the threshold contribution, T_s is the expected location of the threshold as temperature increase above pre-industrial levels, σ is the standard deviation of the normal distribution in the location of the threshold, $T(t)$ is temperature at time t. The threshold as a share of GDP is symmetric for all regions of the model.

For the following runs we fixed $\sigma = 0.05$, which induces a continuous function that is very close to a step in damages (for $T_s = 2.5$ and $d = 0.04$, the damage at $T = 2.45$ is only at 0.0031). The location of the threshold and the maximum damages were varied. For most of the runs, d was at 0.04. The final damage costs that enter the budget equation are:

$$D(i,t) = \left[\Omega(i, T(t)) + d * erf\left(\frac{T(t) - T_c}{\sigma}\right)\right] * GDP(i,t)$$

In order to find the equilibrium in emission strategies in the second stage of the game, both models applied an iterative approach. Emissions of other regions are fixed while non-signatories maximize individual welfare and the coalition maximizes social welfare. We find that individual regions hardly change their abatement strategies when introducing a threshold.

4.2. Stochastic Thresholds

Introducing uncertainty in numerical models is a challenging task. We therefore approximate decision under uncertainty in MICA and WITCH by assuming that in the participation stage it is uncertain whether there exists a threshold in the climate system or not. While the participation decision is taken under uncertainty based on expected utility, the emission strategies are determined under certainty for different states of the world.

After the coalition has formed, we allow for two states of the world in the following way: at some point in time in the future, the true state of the world is known, and the model is run deterministically and with full information from this point on. Prior to this "learning time", we fix the model to a scenario. For this scenario, we chose the "deterministic equivalent", i.e. deterministic decision assuming the expected value of the uncertain parameters was true.[7] We refer to this initial period as the "non-anticipation periods". This procedure allows to approximate decision under uncertainty. Thresholds are implemented the same way as in section 4.1.

So for each coalition, we have computed four scenarios:

1	with threshold	without non-anticipation periods
2	without threshold	without non-anticipation periods
3	with threshold	with non-anticipation periods
4	without threshold	with non-anticipation periods

The expected utility is calculated through combining either scenarios 1 and 2 or scenarios 3 and 4: we sum the utility levels of the two states weighted by the probability of the state occurring.

5. Results

This section describes the results when introducing thresholds in the numerical models MICA and WITCH. We first describe the behavior of regions with respect to emissions and membership in the

[7] It is known that in many settings this "deterministic equivalent" is close to the hedging strategy that results from decision under uncertainty.

agreement when the location of the threshold is known with certainty. The second part reports results when introducing uncertainty about the presence of the threshold.

5.1. Deterministic thresholds

Analytical calculations show that especially deterministic thresholds enhance the scope for cooperation. This section first explores the second stage of the game and describes how coalitions adjust their abatement and whether they keep temperatures below the threshold. In the second part we discuss in how far the membership decision of the first stage of the game is affected by thresholds in the climate system.

5.1.1.Abatement behavior of coalitions

We observe three different kinds of abatement strategies of coalitions in the presence of a threshold:

i. The coalition increases abatement so to keep temperature below the threshold for the entire time-horizon.
ii. The coalition postpones the crossing of the threshold in time to avoid severe damages early on by moderately increasing abatement.
iii. The coalition hardly changes abatement due to the presence of the threshold.

While behavior (i) and (iii) are equivalent to outcomes of the simple analytical model of section 3.1, behavior (ii) induces moderate changes to abatement, which is a dynamic effect that has so far not been considered.

Figure 1 displays the three different types of abatement strategies for the case of the grand coalition of all regions facing different (but certain) locations of the threshold.

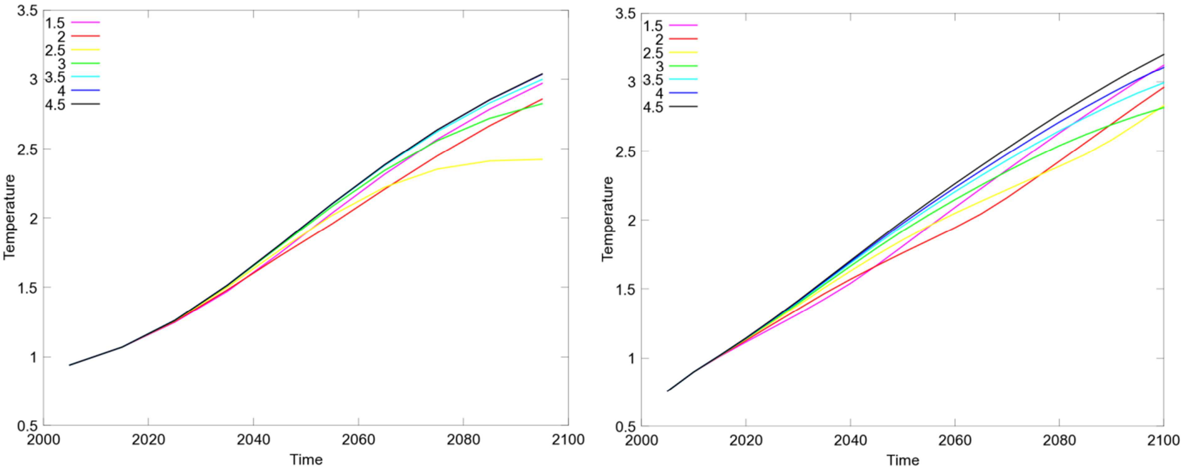

Figure 1: Temperature over time for different locations of the threshold T_S for MICA (left) and WITCH (right), $d = 0.04$ and $\sigma = 0.05$

In MICA, the grand coalition keeps the thresholds for temperatures above or equal to $T_S = 2.5$°C, while for $T_S = 2$°C or lower staying below the threshold is too costly. Still, before the threshold is crossed, we observe a lower temperature compared to the reference: the temperature stays lowest until 2070, i.e. a postponing of crossing the threshold. This is more pronounced at $T_S = 2$°C than at $T_S = 1.5$°C, which has a temperature path hardly different from the reference without a threshold. $T_S = 4$°C is just below the temperature that the Grand Coalition would keep without additional damages. Therefore, the black curve represents the temperature profile the grand coalition of all regions would achieve in the absence of thresholds.

In WITCH, only the second type of behavior is observable. The threshold is crossed at some point for all threshold locations. The main reason can be seen in the high degree of inertia in the energy system resulting in higher abatement costs and less when-flexibility. Also, the shorter time horizon of 2150 makes the long-term welfare costs comparably lower and hence crossing the threshold less detrimental. Still, for lower values of the temperature threshold, the crossing is somewhat delayed while after it is crossed, temperature increases actually accelerate. For T_S =3 and 3.5°C, the temperature increases above the threshold only after 2100 due to decreased emissions by the coalition. In all scenarios, the grand coalition of all regions achieves a later crossing of its threshold by increasing abatement: Until about 2060, the temperature level increases at specific points in time when going from scenario T_S =2°C to $T_S = 4.5$°C. Afterwards, the grand coalition in the T_S =2°C scenario has crossed the threshold and stops additional abatement. The point of ending ambitious mitigation occurs later in time the higher the threshold temperature.

The different regimes of behavior are summarized in figure 2, which displays the temperature in 2100 under different locations of the threshold. For a low location of the threshold T_S, keeping the threshold is too ambitious for the coalition and temperatures in 2100 are high. Increasing T_S, it pays for the coalition to postpone the crossing of the threshold in time (also for the entire time horizon in some cases) and the temperature decreases. For even higher values of the threshold, the coalition will keep the temperature below T_S in 2100 and the temperature increases with the threshold location. These results can be understood when interpreting equation (3): If keeping the threshold is too costly (high abatement costs c or too much abatement effort $\bar{Q}$) in comparison to the damage costs it induces B, abatement will decrease to the level without the presence of a threshold. In the numerical models, this decision is now spread over the entire time-horizon and can be taken for each time-period. Hence, we observe the postponement behavior of coalitions.

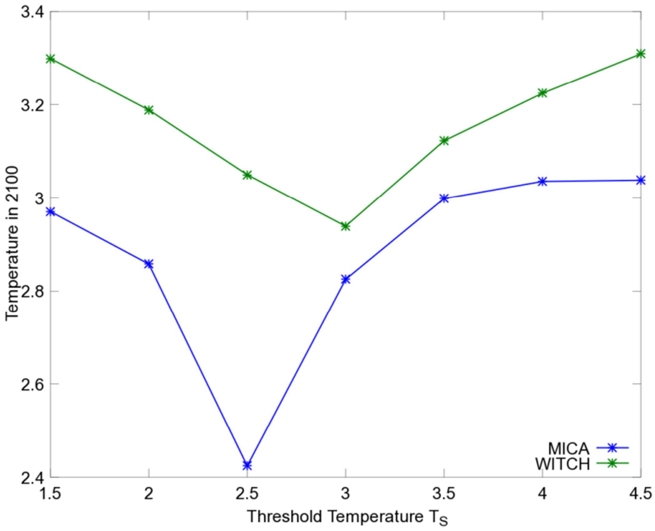

Figure 2: Temperature in 2100 reached by the grand coalition of all regions for different locations of the threshold, $d = 0.04$ and $\sigma = 0.05$

Considering the sub-coalitions to the grand coalition when one region leaves all three different types of behavior are present in MICA and WITCH. Tables 1 and 2 display numbers on how much the remaining signatories of the grand coalition increase or decrease their abatement effort when each region leaves the grand coalition. In table 1 for MICA, the numbers show that for some scenarios the remaining signatories increase their abatement effort when a region leaves a coalition (when numbers are positive: emissions of a group of regions are cumulatively higher when the grand coalition of all regions forms as opposed to when one region leaves). This behavior is contrary to the abatement strategies that were previously described in the literature when continuous damage costs were assumed (see Karp and Simon 2013). A coalition would internalize the climate damages of a region only if it is a signatory to the agreement, hence abatement would be increased as the coalition grows in size. However, if a smaller coalition decides to keep the threshold, temperatures do not change much when free-riding and the coalition actually increases abatement when the number of signatories decreases. On the other hand, for some scenarios emissions are drastically increased when a region leaves the grand coalition, see T_S =2.5°C for all regions besides Russia (RUS) and Japan (JPN). In these scenarios, the sub-coalitions decide not to keep the threshold anymore for a much longer time period than in the grand coalition. Moderate changes in cumulative emissions indicate that behavior (ii) is optimal for some coalitions and abatement is only marginally changed when the coalition becomes smaller (see for example Africa (AFR) for T_S =3°C).

Table 1: Difference in cumulative emissions until 2100 of all regions but the indicated region between the grand coalition run and the sub-coalition when the indicated region leaves (in GtC), for different locations of the threshold in MICA, $d = 0.04$ and $\sigma = 0.05$

$T_S =$	ROW	AFR	LAM	IND	CHN	MEA	OAS	RUS	JPN	USA	EUR
1.5	-5	-61	-12	-52	-28	-11	-23	-5	-2	-18	-16
2	0	-49	-1	-43	-75	-4	-14	-3	1	-13	-12
2.5	-294	-420	-387	-410	-320	-377	-391	6	26	-295	-308
3	46	-29	65	52	144	74	87	3	20	40	28
3.5	24	60	31	13	72	42	57	-1	10	15	11
4	1	-8	-6	-37	-6	1	7	-4	0	-10	-9
4.5	-5	-56	-14	-53	-30	-12	-22	-5	-3	-17	-15

In WITCH, only moderate changes of abatement are observable. Still, the presence of thresholds leads to more drastic changes in temperature for some locations of the threshold (see table 2). In some scenarios abatement efforts of the sub-coalition increase when one region leaves, see for example

T_S =4°C when China leaves. The effects are less pronounced for WITCH compared to MICA, which is due to the more costly abatement and inertias of the energy system as well as the shorter time horizon of the WITCH model. Hence, the abatement behavior of coalitions is much more in accordance to the case of continuous damages.

Table 2: Difference in cumulative emissions until 2100 of all regions but the indicated region between the grand coalition run and the sub-coalition when the indicated region leave (in GtC), for different locations of the threshold in WITCH, $d = 0.04$ and $\sigma = 0.05$

$T_S =$	USA	OLDEURO	NEWEURO	KOSAU	CAJAZ	TE	MENA	SSA	SASIA	CHINA	EASIA	LACA	INDIA
1.5	-69	-77	-8	-23	-33	-14	-31	-104	-45	-37	-58	-19	-107
2	-86	-69	-12	-20	-25	-6	-20	-81	-33	-51	-41	-19	-90
2.5	-72	-64	0	-12	-13	-25	-28	-71	-13	-44	-32	-25	-81
3	-65	-69	-4	-14	-15	-18	-10	-46	-18	-83	-16	-14	-58
3.5	-34	-27	2	-6	-12	25	-1	-37	-5	3	-17	-1	-39
4	-24	-30	2	-6	-16	26	1	-41	-28	21	-19	0	-37
4.5	-19	-23	4	2	-5	31	8	-40	-17	28	-24	4	-39

5.1.2.Stability results

The different abatement behaviors of coalitions induce strategically different incentives to sign the agreement. Free-riding of a certain signatory is greatly enhanced if both the grand coalition and the sub-coalition (when the signatory leaves) keep the threshold over the entire time-horizon: the free-rider is decreasing individual abatement costs upon leaving while damages do not change as there are no temperature changes. If the grand coalition switches from keeping the threshold to hardly increasing abatement upon its presence when a signatory leaves, damages sharply increase upon leaving inducing free-riding to become much less attractive. Both strategic behaviors are already known from the analytical model. Moderate changes to the abatement effort of the coalition, as when switching from keeping the threshold to postponing it in time, will induce moderate changes to damages and are therefore much more unlikely to induce participation of that region.

The different abatement efforts translate into stabilities, displayed in table 3 for MICA: When the temperature changes only marginally when leaving the grand coalition, the stability function is negative as free-riding is very attractive (small numbers in table 1 for differences in emissions). This is different for the case of T_S =2.5°C. Here, most sub-coalitions do not keep the threshold for most of the models time horizon and emissions increase greatly upon leaving for nine regions, giving a positive incentive to stay for all regions besides China (CHN), Middle East and North Africa (MEA), Russia (RUS) and Japan (JPN). The abatement potential of the latter two regions is so small that when leaving, the remaining coalition is still able to keep the threshold at small additional abatement costs, inducing an incentive to leave. For China and Middle East and North Africa, their abatement potentials are large and they incur very high costs inside the grand coalition. Comparing these saved abatement costs to the gains from keeping the threshold results in a positive incentive to leave although the threshold is crossed.

Because changes in abatement efforts are only moderate in WITCH, the value of the stability function is negative for all regions and scenarios, which is shown in figure 3 to be discussed below.

Table 3: Stability function (see equation 1) of the respective region leaving the grand coalition of all regions in MICA, $d = 0.04$ and $\sigma = 0.05$

$T_S =$	ROW	AFR	LAM	IND	CHN	MEA	OAS	RUS	JPN	USA	EUR
1.5	-2.14	-25.50	-10.21	-7.90	-22.98	-10.27	-15.29	-0.29	-4.02	-0.33	-0.27
2	-3.22	-33.61	-14.34	-21.95	-21.10	-15.40	-21.52	-0.38	-5.29	-1.09	-1.06
2.5	0.15	8.53	1.60	40.68	-35.34	-9.46	11.61	-0.53	-7.46	5.02	5.82
3	-2.99	-8.71	-10.32	-39.67	-46.22	-17.46	-30.55	-0.19	-3.45	-1.34	-1.38
3.5	-1.66	-34.04	-5.90	-15.19	-18.53	-9.12	-17.35	-0.10	-2.07	-0.60	-0.70
4	-1.22	-21.29	-4.54	-3.96	-9.79	-6.17	-10.61	-0.06	-1.69	-0.35	-0.45
4.5	-1.16	-13.85	-4.33	-0.69	-8.09	-5.71	-8.40	-0.06	-1.63	-0.31	-0.40

Due to the heterogeneous setting of MICA, the presence of thresholds does not induce stability of an entire coalition (as in the simple analytical model) but internal stability of some regions only. The regions that have an incentive to stay inside the grand coalition of all regions have specific characteristics. First and foremost, the abatement potential of the leaving signatories needs to be large so that keeping the temperature below the threshold becomes costly and unattractive for the coalition when that region leaves: the sub-coalition crosses the threshold temperature earlier inducing high additional damages upon leaving. These regions are pivotal in the sense that their membership is necessary to keep the threshold. Additionally, increased abatement costs need to be valued against the benefits of keeping the threshold for an individual region. If damages are not sufficiently high and individual abatement is too costly in comparison, leaving the coalition can become attractive, even if the threshold is crossed upon leaving. The endogenous interplay between abatement costs and damage costs therefore determines stability in a complex manner. Regions need to be pivotal to keep the threshold but also their individual benefits from keeping the threshold need to be high enough.

For $T_s = 2.5$°C seven out of the eleven regions have a positive incentive to sign the grand coalition agreement. If the surplus of these regions is distributed to the regions that lose from cooperating, stability of the grand coalition could be achieved. We use the method derived in Kornek et al (2014) in order to test if there exist transfers that once implemented realize a positive incentive to sign for all regions.

Table 4 displays for which combinations of location of the threshold and maximum damage costs there exists a transfer mechanism within the grand coalition such that each region has a positive incentive to sign. For three out of these 20 scenarios the surplus of some regions was enough to compensate those regions that lose from cooperation. For T_s <2.5°, already the grand coalition does not keep the threshold and the abatement behavior goes back to the normal one with continuous damages, inducing all regions to have an incentive to leave the grand coalition and no scope for transfers to enhance cooperation. For T_s >2.5°C, too many sub-coalitions also keep the threshold, at least partially, such that there are too view regions with an incentive to sign the agreement.

T_s =2.5°C is the only threshold temperature that induces sufficiently many regions to have a positive incentive to sign the agreement that compensation of the other regions is possible. Still, a transfer scheme does not exist for all damage costs: if the threshold damage increases to d =0.045, more sub-coalitions to the grand coalition act strongly upon the presence of the threshold, making free-riding more attractive and impeding potential internal stability.

Table 4: Indication if there exists a transfer mechanism inside the grand coalition of all regions such that every region has a positive incentive to sign the agreement, for different values of threshold location T_S and maximum damage costs d in MICA ($\sigma = 0.05$): "1" indicates that there exists a transfer mechanism

T_s\\d	0.03	0.035	0.04	0.45

2.3	0	0	0	0
2.4	0	0	0	0
2.5	1	1	1	0
2.6	0	0	0	0
2.7	0	0	0	0

In conclusion, the presence of the threshold enhances the prospect for cooperation, but only for very specific parameter ranges. We identified that threshold temperatures of few degrees above the pre-industrial level critically influence abatement of the grand coalition. In addition, countries with large abatement potential are more likely to become to have an incentive to sign in the presence of thresholds since they assess that a coalition without their membership will not keep temperature below the threshold. However, still then the trade-off between abatement costs and avoided damages might impede cooperation. In addition, we identified crucial differences between the models MICA and WITCH. In MICA, changing abatement is much more flexible due to the simple representation of abatement opportunities. In WITCH, abating is connected to large inertias in the energy system and higher abatement costs. The time horizon is also shorter which decreases benefits from lower temperatures. Hence, moderate changes to abatement are more likely in WITCH, which leads to the typical free-riding behavior that was described in previous literature.

5.2. Uncertain thresholds

In order to exemplarily test the effects of uncertainty in MICA and WITCH, we allowed for two states of the world: one without a threshold and one with ($d = 0.04$, $\sigma = 0.05$, in MICA: T_S =2.5°, in WITCH: T_S =3°) with equal probability. For the set of scenarios with a non-anticipation period, emissions were fixed for 30 years. If decisions are based on expected utility, both for abatement decisions with and without anticipation period, stability is impeded in MICA and WITCH. The reduced prospect for cooperation has a number of reasons, and we use calculations for the grand coalition to illustrate these points in figure 3. First and foremost, the possibility of a world without a climate damage threshold induces has an adverse effect on expected value of the stability function for every region since free-riding to the grand coalition is highly attractive in the absence of thresholds. This offsets any positive effect that the presence of a threshold in the other state of the world might have, and therefore lowers the values of the stability function by construction. Second, when including a non-anticipation period, staying below the threshold for the grand coalition becomes more costly (not shown) because the abatement during the non-anticipation time is suboptimal and hence reduces the utility of signatories. However, the utility as a free-rider is hardly affected as the threshold was not kept by almost all sub-coalitions without the non-anticipation period anyway. Hence, prospects for cooperation worsen in general when uncertainty is introduced. However, we also observe some positive values for the stability function for some regions in MICA if thresholds are uncertain, see USA and Europe (EUR) in figure 3. For these regions, the benefits from keeping the threshold if it exists are so large that they compensate the possible losses when the threshold does not exist. Hence, the presence of the uncertain threshold may still increase the scope for cooperation.

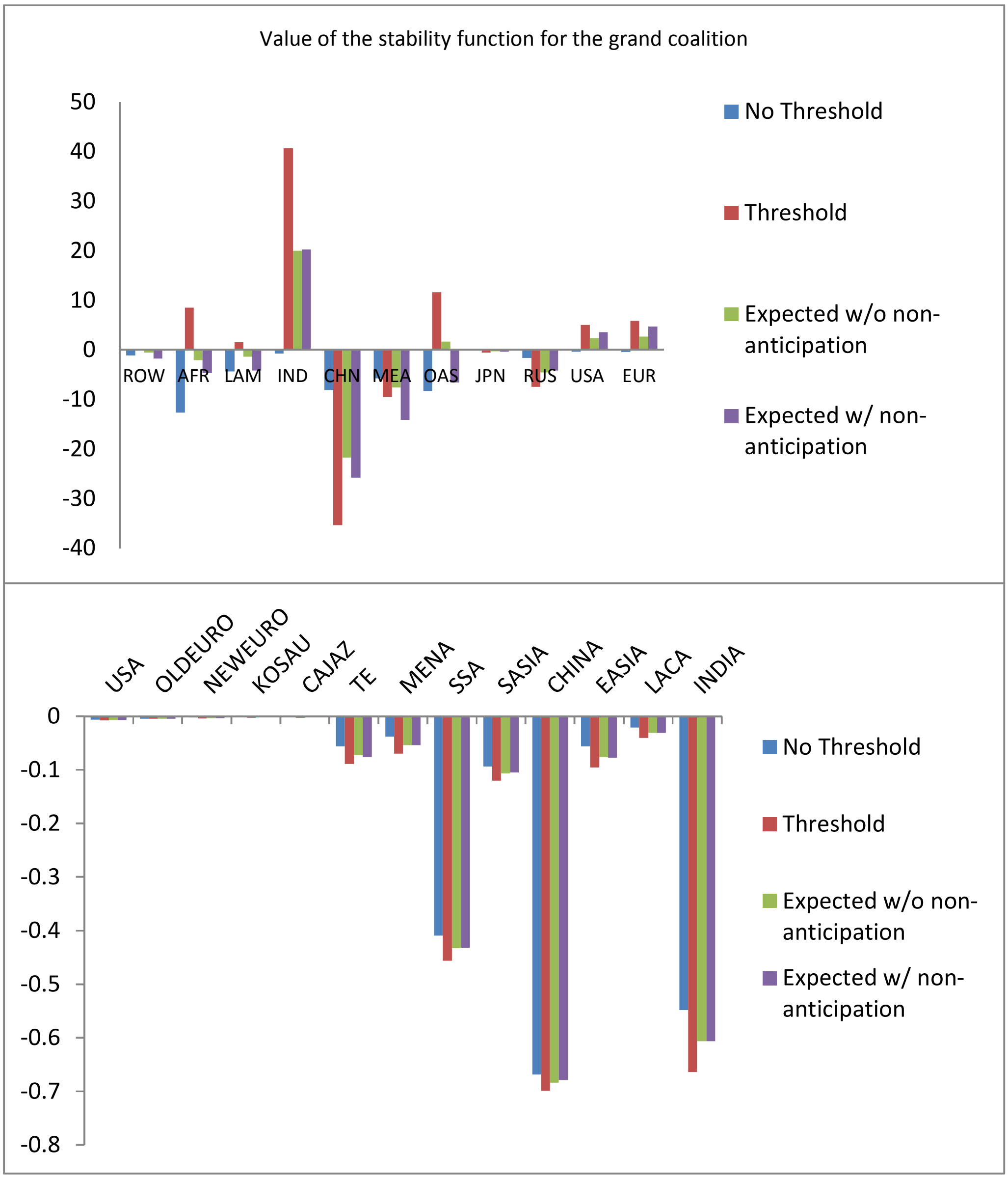

Figure 3: Value of the stability function (see equation 1) for four different scenarios: Without the presence of a threshold (blue); with a presence of a threshold (red); Expected Utility without a non-anticipation period (green); Expected Utility with a non-anticipation period (purple); MICA in the upper figure with , $\boldsymbol{T_s = 2.5°\mathrm{C}, \mathrm{d} = 0.04,\ \sigma = 0.05}$, WITCH in the lower figure with , $\boldsymbol{\mathrm{T_S} = 3.0°\mathrm{C}, \mathrm{d} = 0.04, \sigma = 0.05}$

The two states of the world are assumed to occur with equal probability. This includes the possibility of no threshold existing at all, arguably this assumption reflects very large uncertainty about the threshold.

Further investigations and more levels of uncertainty are necessary. Our model nevertheless illustrates the difficulties that arise once thresholds are uncertain.

6. Conclusions

Climate change remains a daunting challenge for the international community. A large body of academic literature has assessed that the public good nature of abating greenhouse gas emissions impedes cooperation because countries find themselves in situation resembling a classical “Prisoner’s Dilemma”. Recent literature has however discussed how the stability of coalitions and environmental quality is enhanced when thresholds in the damage costs are introduced in the analysis. However, this result depends on the shape of the threshold considered.

The numerical analyses with the models MICA and WITCH show that the socially optimal abatement – when all regions cooperate – keeps temperatures below a threshold of moderate warming (approximately at 2.5 °C or higher) and of sufficiently large damage costs (several percentage points of national product). Otherwise, abatement costs are too high compared to the damage costs so that keeping the threshold is not Pareto-efficient. When one region defects from the grand coalition of all regions, it may be optimal for the remaining regions to either keep the threshold for the entire time horizon, stay below the threshold temperature only temporarily or drop abatement to the level that would be optimal without the presence of the threshold.

When a member leaves, the reaction of coalitions can therefore be contrary to what has been described in previous literature. If the sub-coalition finds it optimal to keep the threshold, abatement actually increases when the size of the coalition becomes smaller. The leaving region has a high incentive to free-ride because the damage costs do not increase but abatement costs decrease significantly. Hence, cooperation is impeded in this case.

If, on the other hand, the sub-coalition decreases abatement such that the threshold is not kept anymore, damage costs increase sharply for the free-riding region. We emphasize the presence of these pivotal regions whose mitigation potential is critical to keep temperatures below the threshold. If the decrease in abatement costs upon leaving is not too high compared to the increase in damage costs, pivotal regions may find it optimal to sign the agreement. In MICA, the threshold at $T_S = 2.5$ °C induces seven out of eleven world regions to have a positive incentive to join the grand coalition when the threshold induces a percentage loss of GDP of four percent. The presence of thresholds may therefore enhance the prospects for cooperation. We show that the stability of the entire coalition can be achieved if regions with a positive incentive to sign compensate the other regions for their mitigation effort.

Compared to MICA, the coalitions in WITCH mostly adjust abatement in a more moderate manner in the presence of thresholds and for different regions leaving the grand coalition of all regions. The WITCH model therefore assesses less scope for cooperation mostly due to different representations of dynamic abatement. Higher abatement costs in WITCH and the shorter time-horizon results in less benefits from keeping the threshold.

In a last exercise, we introduce uncertainty about the presence of a threshold. We confirm the literature in showing that the scope for cooperation is worsened. However, some regions may still have a positive incentive to sign in the presence of an uncertain threshold as opposed to the absence of a threshold. The analysis is exemplary and future research is going to broaden the parameter space on uncertainty.

7. References

Scott Barrett. Self-enforcing international environmental agreements. Oxford Economic Papers, 46:878–894, 1994.

S. Barrett. Environment and Statecraft: The Strategy of Environmental Treaty-Making. Oxford University Press, 2003.

S. Barrett and A. Dannenberg. Climate negotiations under scientific uncertainty. PNAS 109(43): 17372–17376. 2012

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

V. Bosetti, C. Carraro, M. Galeotti, E. Massetti, and M. Tavoni. WITCH: A World Induced Technical Change Hybrid model. The Energy Journal, Special Issue Hybrid Modelling of Energy Environment Policies: Reconciling Bottom-up and Top-down(27):13–38, 2006.

P. Chander, and H. Tulkens. A core-theoretic solution for the design of cooperative agreements on transfrontier pollution. International Tax and Public Finance, 2:279–93, 1995.

C. d'Aspremont and J. J. Gabszewicz. New developments in the analysis of market structures, chapter On the stability of collusion, pages 243–64. Macmillan, New York, 1986.

M. Finus. Game Theoretic Research on the Design of International Environmental Agreements: Insights, Critical Remarks, and Future Challenges. International Review of Environmental and Resource Economics, 2:29–67, 2008.

M. Hoel. International Environment Conventions: The Case of Uniform Reductions of Emissions. Environmental and Resource Economics 2:141-159, 1992.

L. Karp and L. Simon. Participation games and international environmental agreements: A non-parametric model. Journal of Environmental Economics and Management, 65:326–344, 2013.

U. Kornek, K. Lessmann, H. Tulkens (2014), Transferableand Non Transferable Utility Implementations of Coalitional Stability in Integrated Assessment Models. CORE Discussion Paper 35, 2014.

T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H.-J. Schellnhuber. Tipping elements in the Earth's climate system. PNAS, 105(6):1786–1793, 2008.

K. Lessmann, R. Marschinski, and O. Edenhofer. The effects of tariffs on coalition formation in a dynamic global warming game. Economic Modelling, 26(3):641–649, 2009.

K. Lessmann, U. Kornek, V. Bosetti, R. Dellink, H.-P. Weikard, J. Emmerling, M. Nagashima, and Z. Yang. The stability and effectiveness of climate coalitions: A comparative analysis of multiple integrated assessment models. Fondazione ENI Enrico Mattei (FEEM) Nota di Lavoro 5.2014, 2014.

W. Nordhaus. Managing the Global Commons. MIT Press, 1994.

W. D. Nordhaus, and Z. Yang. A regional dynamic general-equilibrium model of alternative climate-change strategies. The American Economic Review, 86(4):741–765, 1996.

J. Rogelj, C. Chen, J. Nabel, K. Macey, W. Hare, M. Schaeffer, K. Markmann, N. Höhne, K. Krogh Andersen, and M. Meinshausen. Analysis of the Copenhagen Accord pledges and its global climate impacts- a snapshot of dissonant ambitions. Environmental Research Letters, 5:034013 (9pp), 2010.

H.-P. Weikard. CARTEL STABILITY UNDER AN OPTIMAL SHARING RULE. The Manchester School, 77(5):1463–6786, 2009.

8. Appendix

For the social optimum not to be a Nash-equilibrium, the following condition needs to hold:

$$N\frac{B}{c} - \bar{Q}\Delta Q - \frac{\Delta Q^2}{2} < 0. \qquad \text{A.1}$$

For a coalition of size k to keep the threshold, two conditions are sufficient: (i) the utility at that point needs to be greater than the Nash-equilibrium utility (which is zero) and (ii) the derivative of the joint utility with respect to abatement needs to be non-negative in the range $\left[\bar{Q} - \frac{\Delta Q}{2}, \bar{Q} + \frac{\Delta Q}{2}\right]$ (utility is maximal when $q = \frac{1}{k}(\bar{Q} + \frac{\Delta Q}{2})$ for coalition signatories). Assuming that the free-riders do not abate at all, this gives:

$$\pi_m\left(q^m = \frac{1}{k}\left(\bar{Q} + \frac{\Delta Q}{2}\right), q^{nm} = 0\right) > 0 \Rightarrow k^2\frac{B}{c} - \frac{1}{2}\bar{Q}^2 - \bar{Q}\frac{\Delta Q}{2} + \frac{\Delta Q^2}{8} \geq 0$$

A.2

$$\frac{\partial \pi_m}{\partial q_i}\Big|_{q^m = \frac{1}{k}\left(\bar{Q} + \frac{\Delta Q}{2}\right)} > 0 \Rightarrow k^2\frac{B}{c} - \bar{Q}\Delta Q - \frac{\Delta Q^2}{2} \geq 0.$$

A.3

With these conditions, the signatory utility to a coalition of size k is just positive. If a coalition of size $(k-1)$ would switch to zero abatement, the coalition of size k is internally stable and therefore also externally stable (utility of a joining signatory would decrease below B). For a coalition of size $(k-1)$ to want to leave the threshold, the derivative of the joint utility with respect to abatement needs to be negative at the threshold. Again, assuming that free-riders do nothing, this leads to the following condition:

$$(k-1)^2\frac{B}{c} - \bar{Q}\Delta Q - \frac{\Delta Q^2}{2} < 0. \qquad \text{A.4}$$

This would mean that the coalition would enter the region of continuous change in benefits. In this region, the coalition maximizes joint utility when setting $q^m = (k-1)\frac{B}{c}\frac{1}{\Delta Q}$. Non-signatories then also have an incentive to increase abatement in the dominant strategy (so no negotiation needed): $q^{nm} = \frac{B}{c}\frac{1}{\Delta Q}$. For the coalition to prefer a zero-abatement strategy, the following condition needs to hold:

$$\frac{B^2}{c\Delta Q^2}[(k-1)^2+(N-k+1)]-\frac{B}{\Delta Q}\left(\bar{Q}-\frac{\Delta Q}{2}\right)-\frac{1}{2}\frac{B^2}{c}\frac{1}{\Delta Q^2}(k-1)^2<0 \qquad \text{A.5}$$

Finally, in order for equation A.5 to make sense, the cumulative abatement of the coalition and the non-signatories needs to below the upper threshold bound:

$$\frac{B}{c\Delta Q}[(k-1)^2+(N-k+1)]<\bar{Q}+\frac{\Delta Q}{2} \qquad \text{A.6}$$

The set of inequalities A.1-A.6 define a stable coalition size k^*if fulfilled. A numerical check confirmed that a solution exists.