Climate thresholds and heterogeneous regions: implications for coalition formation*

Johannes Emmerling† Ulrike Kornek‡ Valentina Bosetti§ Kai Lessmann¶

January 17, 2020

Abstract

The threat of climate catastrophes has been shown to radically change optimal climate policy and prospects for international climate agreements. We characterize the strategic behavior in emissions mitigation and agreement participation with a potential climate catastrophe happening at a temperature threshold. Players are heterogeneous in a conceptual and two numerical models. We confirm that thresholds can induce large, stable coalitions. The relationship between the location of the threshold and the potential for cooperation is non-linear, with the highest potential for cooperation at intermediate temperature thresholds located between 2.5 and 3 degrees of global warming. We find that some regions such as Europe, the USA and China are often pivotal to keeping the threshold because the rest of the world abandons ambitious mitigation and the threshold is crossed without their participation. As a result, their incentives to cooperate can be amplified at the threshold. This behavior critically depends on the characteristics of the threshold as well as the numerical model structure. Conversely, non-pivotal regions are more likely to free-ride as the threshold inverts the strategic response of the remaining coalition. Moreover, we find that our results depend on which equilibrium concepts is applied to analyze coalition formation as well as the introduction of uncertainty about the threshold.

Keywords: tipping points, international environmental agreements, climate change

JEL Classification: C72, D62, H41, Q54

Please cite as: Emmerling, J., U. Kornek, V. Bosetti, K. Lessmann (2020): Climate thresholds and heterogeneous regions: implications for coalition formation. The Review of International Organizations, doi:10.1007/s11558-019-09370-0.

*We thank participants at the FEEM Workshop on Public Goods 2014, the 20th Coalition Theory Network Workshop in Venice, the 7th Atlantic Workshop on Energy and Environmental Economics in Atoxa, two anonymous referees, Alessandro Tavoni, Henry Tulkens, and Philippe Colo for very helpful comments. VB would like to acknowledge financial support from the ERC grant agreement n° 336703 (RISICO). KL gratefully acknowledges financial support by the Federal Ministry of Education and Research (BMBF) program "Global Change 5+1" as part of the grant agreement 01LN1703A (FINFAIL). An earlier version of this paper has been circulating under the title "The catastrophe smile - The effect of climate thresholds on coalition formation".

†RFF-CMCC European Institute on Economics and the Environment (EIEE), Centro Euro-Mediterraneo sui Cambiamenti Climatici, Via Bergognone, 34, 20144 Milano, Italy, E-Mail: johannes.emmerling@eiee.org

‡Mercator Research Institute on Global Commons and Climate Change (MCC), Torgauer Str. 12-15, 10829 Berlin, Germany and Potsdam Institute for Climate Impact Research (PIK)

§Department of Economis Università Bocconi, via Roentgen, 1, 20136 Milan, Italy and RFF-CMCC European Institute on Economics and the Environment (EIEE), Centro Euro-Mediterraneo sui Cambiamenti Climatici

¶Potsdam Institute for Climate Impact Research (PIK), PO Box 60 12 03, D-14412 Potsdam, Germany

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 (notably the Copenhagen pledges and the Paris agreement) so far fall short of their own ambition (Rogelj et al. 2010, Rogelj et al. 2016). The global public good nature of mitigation impedes comprehensive cooperation because free-riding on other countries is possible while individual costs are avoided (Barrett 2003). The design of international environmental agreements aims to overcome this incentive problem.

Early coalition formation literature studies the incentives of countries to sign an international climate agreement based on the trade-off between costs of emission reductions versus damage costs from emissions (Hoel 1992, Barrett 1994). An important result from this literature is that large coalitions are only stable if they do not need to achieve much. Finus (2008) and Benchekroun and Van Long (2012) highlighted how the agreement's design may be modified to improve participation and its environmental effectiveness. One of the most important features favoring climate cooperation is heterogeneity or differences with respect to costs of mitigation and the associated damages across countries. Incorporating transfers between regions allows larger coalitions to become stable (Nagashima et al. 2009, Weikard 2009, Lessmann et al. 2015).

Other properties of climate change, beyond its public good character, can also be explored with respect to their implications for cooperation. Notably, the existence of potential catastrophic climate damages has received great attention in the impact literature. Catastrophic impacts, threshold damages and tipping points and their role for optimal management of environmental systems have been at the core of a large strand of environmental literature (see, e.g., Muradian, 2001; Brozovic and Schlenker, 2011). The majority of the climate coalition formation literature only considers continuous damages from greenhouse gas emissions but recent studies have emphasized the role of such 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 compared to the costs of mitigation. While individual countries are not able to keep the threshold by themselves and the country-specific losses from crossing it are relatively low, an international agreement is a means for countries to coordinate on the social optimum. This solution presents a non-cooperative Nash-equilibrium of the game in emissions strategies. The agreement thus serves as a means of coordination if a threshold of sufficient characteristics is present.

Barrett and Dannenberg (2012) and Barrett (2013) highlighted that uncertainty about the location of the threshold may again reverse the implications of thresholds for coordination. If the exact amount of emissions to avoid crossing of the climate threshold is unknown, the point of reference for coordination vanishes. Still, there exists a range of parameter values for which the problem of climate change may still be a coordination game. This problem has been further analyzed theoretically and in an experimental context in Schmidt (2017) and Iris and Tavoni (2016). In the context of a renewable resource, Miller and Nkuiya (2016) also studied the possibility of cooperation under (uncertain) thresholds showing similar ambiguous results. Polasky, de Zeeuw, and Wagner (2011) showed how tipping points induce a precautionary optimal policy. In a game theoretic context, and based on stochastic-dynamic model of tipping points, Sakamoto (2014) and Diekert (2017) showed how a

tipping point can alleviate cooperation. This result builds on the qualitatively similar shallow-lake problem (Mäler, Xepapadeas and de Zeeuw, 2003), based on non-linear dynamics of a common property resource. However, this literature has focused on the stochastic properties, learning, and notably symmetric equilibria in such games. In our application, we focus on the role of heterogeneity, which in the global context of climate change is a crucial feature to consider.

This makes the characteristics of climate thresholds crucial for their role to coordinate on admissible emissions. 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 above 1980-1999 levels – the arctic summer ice at 0.5 to 2°C or the Greenland ice sheet at 1 to 2°C temperature increase in the 21st century. Uncertainties remain large with respect to the exact location of the threshold and the actual impact that the crossing of the threshold would have on economic and social systems. Based on a large expert elicitation survey, Kriegler et al. (2009) provide probabilistic estimates of the distribution of these parameters. In terms of their role for the globally optimal climate policy, these tipping points have recently been integrated in numerical models to assess problems such as the optimal mitigation policy (Cai et al., 2016; Lemoine and Traeger, 2016, Lontzek et al., 2015, Tsur and Zemel, 2016) and that of investment in Solar Radiation Management (Heutel et al., 2016).

In this paper, we analyze the effect of climate thresholds on global cooperation. We show the basic mechanics in a simple conceptual model and rely on two numerical climate coalition formation models to test how the effects play out in a real world calibration (MICA, cf. Lessmann et al. 2009, 2015; Kornek et al. 2017; and WITCH, cf. Bosetti et al., 2006; Emmerling et al., 2016). Our approach allows us to contribute in three respects. First, each world region's emission reduction costs and damages are empirically calibrated, allowing for realistic differences between world regions. Second, the characteristics of the threshold can be studied numerically based on the empirical foundation of the climate system and its impact on GDP and consumption. Third, our study extends the analytical literature from a static to a dynamic setting, in which the costs of avoiding emissions 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.

In our analysis, we first outline an analytical model of catastrophic damages under heterogeneity, and then introduce climate thresholds in both numerical models and explore to what extent different locations and economic impacts of the threshold influence optimal emissions strategies in the social optimum. We find that the socially optimal emissions strategy depends on the location of the threshold temperature and takes four different forms: (1) At very high threshold values, the threshold becomes nonbinding with no effect on emission strategies; (2) For lower threshold values, the coalition avoids the catastrophic damages, staying below the threshold temperature for at all times; (3) At still lower threshold values, the threshold temperature is eventually exceeded but at a later date, postponed compared to the absence of threshold damages; (4) At very low threshold values, the coalition resigns to abate as would be optimal in a scenario without the existence of the threshold. These different emissions strategies result in a "catastrophe smile" with cumulative emissions going from high to low and back up to high. We find that the coalition of all countries avoids the threshold when its location is in the range of 2.5 to 3.5°C in MICA and for a threshold location of 2.5°C in WITCH.

Then we study the incentive for single regions to leave the grand coalition. Higher emissions by the remaining subcoalitions are the usual response to defection and a potential deterrent to free-riding. With threshold damages, this response changes with the characteristics of the defector. When the additional emissions reduction required by the coalitions to keep temperatures below the threshold despite the defection is sufficiently small, the defector is not pivotal and the remaining coalition still keeps the threshold. This creates a very strong incentive to defect because the defection does not have the usual consequences in warming and associated climate change damages. Contrary to the regular behavior of coalitions when damage costs are continuous, we show that the presence of thresholds can imply an increase in a coalition's ambition as a reaction to the free-riding of one region. This type of behavior prevails in both models when the threshold temperature is between 3 to 3.5°C.

Conversely, when it is prohibitively costly for the remaining coalition to compensate the deviation of a defector, the deviating region becomes pivotal to avoiding the threshold damage. Typically, pivotal regions have great mitigation potential, so that without them the threshold becomes unattainable. In MICA, this is the case for 8 out of 11 regions when the threshold temperature is 2.5°C. The remaining coalition only compensates to keep temperatures under 2.5°C when the regions Russia, Japan and Rest of the World drop out. The high impact of crossing the threshold is then likely to deter the pivotal regions from leaving the coalitions. Indeed, at a threshold temperature of 2.5°C the regions Europe, USA, India, Latin America, Other-Asian-Countries and Africa have an incentive to participate in the grand coalition while they lack this incentive if there is no climate threshold. On the other hand, China and the region of Middle-Eastern-Countries lack an incentive to participate in the grand coalition even though the threshold temperature is crossed when they leave. Being pivotal to avoiding the threshold is therefore not enough to induce participation. For China and the Middle East the costs of mitigation so that the temperature stays below 2.5°C are too large to outweigh the benefits, even though they are substantial. For WITCH, we observe the same strategic behaviors qualitatively but the effects are less pronounced than in MICA. Hence, the quantification of the behavior depends on the model structure and a positive effect on stability is only observable in the simpler of the two numerical models with a longer time horizon.

Our findings on the incentives of pivotal and non-pivotal players echo an insight from the literature with symmetric players and possibly uncertain thresholds in Barrett and Dannenberg (2012). A player finds it unattractive to defect from the social optimum when the threshold is crossed upon defection both in Barrett and Dannenberg and our contribution (if benefits outweigh the costs of mitigation). In Barrett and Dannenberg (2012), emissions of the defecting player rise but remain constant for other players. If the location of the threshold is known with certainty, higher emissions induce the crossing of the threshold. In our analysis, when a player leaves the grand coalition her emissions increase and the remaining coalition adjusts its emissions as a response. The threshold is abandoned only if the defecting player is pivotal.[1] The two analyses however differ when players have an incentive to defect. In Barrett and Dannenberg (2012), players find it attractive to defect when introducing uncertainty about the location of the threshold. As before, a defecting player emits more while the remaining

[1] Note that in his analysis of cooperation and catastrophic damages, Barrett (2013) makes a different assumption about the behavior of the remaining coalition: the coalition acts as a Stackelberg leader in emission choices, anticipating the emission choice of the defecting player. In his model, a leaving player will actually reduce his emissions, by force of the coalition.

players stick to their emission strategies. With an uncertain threshold, this defection merely raises the probability of crossing the threshold. The catastrophe does not necessarily materialize. In our contribution, non-pivotal players find it attractive to leave the coalition because the threshold is not crossed upon defection. The catastrophe does not materialize with certainty because the remaining coalition decreases its emissions to avoid the threshold.[2]

Therefore, while the presence of thresholds has potential to foster cooperation, the asymmetry of regions being – or not being – pivotal to avoiding the threshold calls for transfers to redistribute the gains of cooperation within the cooperation. We show that transfer schemes exist so that the grand coalition can be sustained as a stable agreement. The threshold location for which this occurs is in the range of $2.5°C$ and damage costs of a few percentage points of GDP in MICA. This is the location where the coalition of all countries finds it just optimal to avoid the threshold while for lower threshold temperatures this is not the case. In WITCH the same qualitative behavior can be observed but there is less scope for cooperation mostly due to different representations of dynamic emission reduction possibilities and inertia in the energy system resulting in costly changes of mitigation options. In an extension, we also test in how far our positive results for certain thresholds carry over when there is uncertainty about the location of the threshold. Confirming the literature, we find that the scope for cooperation is significantly reduced when introducing uncertainty.

This paper 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 linking threshold damages to socially optimum emissions mitigation and the incentives to free-ride. The implementation and corresponding analyses of behavior at the threshold are reported in Sections 4 and 5. Section 6 concludes.

2 An analytical coalition formation model with thresholds

We study the stability of the grand coalition of all regions, denoted G, following the predominant approach of modeling 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, regions choose economic strategies that determine the emission 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 (similar to the Partial Agreement Nash Equilibrium of Chander and Tulkens, 1995).[3]

Formally, the free-riding incentive can be assessed by studying the stability function, φ_i, which is the difference in utility $\pi_i(S)$ of a region i when being a signatory to the agreement of coalition S and being a non-signatory to the remaining coalition $S \setminus i$:

$$\varphi_i = \pi_i(S) - \pi_i(S \setminus i) \tag{1}$$

[2]We thank an anonymous reviewer for pointing us to the similarities in strategic behavior at the threshold of symmetric players facing uncertainty and heterogeneous players as in this study.

[3]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.

If the stability function is positive, $\varphi_i \geq 0$, for all regions, all regions have an incentive to sign the agreement. If the stability function is negative for some regions, these regions have an incentive to leave and free-ride on the coalition S.

In some cases, the free-riding incentive can be positive for some regions while other regions lack 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.

Now, 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 that shows 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.

Consider N heterogeneous regions interacting via a global public good. Benefits from abatement follow a step function while abatement costs are assumed to be quadratic. Moreover, we consider two periods and a stock pollutant as an approximation of the dynamics of the numerical models. For simplicity, we assume equal mitigation and impact functions over time. These assumptions lead to the following utility function:

$$\pi_i = -\frac{1}{2\alpha_i}[(\varepsilon_{i1} - e_{i1})^2 + \beta(\varepsilon_{i2} - e_{i2})^2] - \delta_{i\sum_j e_{j1} > E_T} - \beta\delta_{i\sum_j (e_{j1}+e_{j2}) > E_T} \tag{2}$$

Here, α_i is the inverse of the slope of abatement costs, ε_{it} are unregulated or baseline emissions, e_{it} are actual emissions strategies, δ_i is the damage if threshold crossed, and E_T is the location of threshold. To make our main points in the most tractable way, we can further simplify by abstracting from discounting ($\beta = 1$) and assume the periods exhibit equal baseline emissions ($\varepsilon_{i1} = \varepsilon_{i2}$), and we denote by $E_B = \sum_j \varepsilon_{j1} + \varepsilon_{j2}$ the cumulative baseline emissions. That is, $E_B - E_T$ gives a measure of how much global mitigation is needed to stay below the threshold forever.

In order to restrict the space of potential equilibria, we impose in the following two conditions that can be interpreted as simple cost optimality conditions, namely intertemporal and inter-regional cost optimality. That is, we assume that

a) regions minimize intertemporal total mitigation costs and

b) regions within a coalition distribute mitigation effort such that marginal mitigation costs are equalized

Both assumptions seem reasonable in the context of international agreement without the possibility of transfers. Taken together, we now show that the grand coalition has four different strategies partitioning the space of the location threshold.

Proposition 1. *Depending on the location of the threshold E_T, there are four types of equilibria for the grand coalition regarding the attainment of the threshold :*

1) $E_T > E_B : e_{it} = \varepsilon_{it} \forall i, t$: Nonbinding

2) $E_B \geq E_T > E_B - \sqrt{8\sum_j \delta_j \sum_j \alpha_j}$: $\sum_j e_j = E_T/2$: Avoidance forever

3) $E_B - \sqrt{8\sum_j \delta_j \sum_j \alpha_j} \geq E_T > \frac{E_B}{2} - \sqrt{2\sum_j \delta_j \sum_j \alpha_j}$: $\sum_j e_{j1} = E_T$,$e_{i2} = \varepsilon_{i2} \forall i$: Postponement (avoidance only in period one)

4) $E_T < \frac{E_B}{2} - \sqrt{2\sum_j \delta_j \sum_j \alpha_j}$: $e_{it} = \varepsilon_{it} \forall i,t$: Resignation

Proof. It is easy to see that the threshold is non-binding if unregulated emissions are lower than the location of the threshold ($E_T > E_B$). When the threshold is binding, the coalition may choose to avoid it. When the emissions are below the threshold in both periods, emission strategies should maximize payoffs within the coalition over time and across regions: Based on cost optimality it is easy to show that (a) intertemporal cost optimality implies $e_{i1} = e_{i2} \equiv e_i$ for all countries. Moreover, within any coalition and without transfers, (b) costs are minimized across regions for equal marginal costs: $\frac{\partial \pi_i}{\partial e_i} = 1/\alpha_i(\varepsilon_i - e_i) = p \forall i$. Since π_i is increasing in e_{it}, cumulative emissions will be just at threshold location: $2\sum_j e_j = E_T$. Solving for the implicit equalized marginal cost or price p, we hence find $p = \frac{E_B - E_T}{2\sum_j \alpha_j}$. The resulting emission strategies are optimal for the grand coalition if they lead to mitigation costs that are lower than avoided damage costs: $2\sum_j \delta_j \geq 2\sum_j \frac{1}{2\alpha_j}(\varepsilon_j - e_j)^2$. Based on the optimality condition of equalized marginal costs, we find each region's mitigation effort as $(\varepsilon_i - e_i) = \alpha_i p$ and using the expression for p and substituting into the condition for maintaining the threshold we find the right-hand side of the condition in (2). If it is not optimal to keep the threshold in both periods, the coalition may still find it optimal to postpone crossing the threshold to the second period. In this case, emissions in the second period are equal to the baseline $e_{i2} = \varepsilon_{i2}$, since π_i is increasing in e_{i2}. In the first period, marginal costs are equalized across regions: $\frac{\partial \pi_i}{\partial e_{i1}} = 1/\alpha_i(\varepsilon_{i1} - e_{i1}) = p' \forall i$. Aggregate emissions in the first period equal the threshold location: $\sum_j e_j = E_T$. Hence, applying the same computations as for the last case, but only for the first period, we find $\sum_j \delta_j \geq \sum_j \frac{1}{2\alpha_j}(\varepsilon_j - e_j)^2$, which yields the right inequality of condition (3). Case (4) is just the opposite case of (3). □

This result shows that the grand coalition has four different strategies in light of the threshold depending on its location. Based on condition (2), it is clear that crossing the threshold can be avoided if damages δ_j are high and/or mitigation costs are low (α_j, the inverse of marginal abatement costs, is high). When threshold damages occur at very low emission levels E_T, the (quadratic) abatement costs outweigh the benefit of staying below E_T, and emissions remain at their baseline level (resignation). Emissions in the first period are just at the threshold location in case postponement is optimal. Only when the threshold location is sufficiently large are costs of avoiding the threshold in both periods low enough to justify the ambitious emission reductions in both periods. Lastly, the coalition falls back to unregulated emissions if the threshold is nonbinding. This relationship interestingly resembles the finding e.g., of Brozovic and Schlenker (2011), who find a similar non-monotonic relationship between the location about an unknown threshold location and precautionary behavior.

Starting from the grand coalition we evaluate now what happens if one player i leaves, that is, the formation of a subcoalition S of size $N-1$. The distinction whether the threshold is exceeded in both periods or one period only does not generate additional insights, hence for the stability analysis, we only consider the cases where the threshold is kept in both periods, or never.[4] Then, the subcoalition

[4] This assumes that the defector falls back to its baseline emissions. In principle, there are many equilibria in emission strategies here, but characterizing them analytically is beyond the scope of this paper.

will keep the threshold in both periods if $\sum_{j=1}^{N-1} \delta_j \geq \frac{(E_B - E_T)^2}{8 \sum_{j=1}^{N-1} \alpha_j}$. That is, it is more likely that the subcoalition keeps the threshold if δ_i is small, that is, the leaving player i suffers small impacts from crossing the threshold, and/or if α_i is small, i.e., the leaving player has relatively high mitigation costs. Intuitively, such countries do not contribute much abatement to a coalition (because of high costs) and contribute little to the necessity to keep the threshold (because of small damages). Hence their defection is a relatively small loss. If the defecting player is not pivotal to the coalition in the sense that her defection does not affect the coalition's decision to keep the threshold, then there is no incentive for it to stay, as in this case the incurred damages remain unchanged.

However, this changes if the defecting player (k) is "pivotal" to keeping the threshold, i.e., the subcoalition reconsiders not to keep the threshold. Based on the proof of Proposition 1, we know that in the grand coalition N, player k's mitigation effort is $(\varepsilon_k - e_k) = \alpha_k p$ and hence its stability function can be computed as $\varphi_k(N) = -\frac{1}{\alpha_k} (\alpha_k p)^2 - (-2\delta_k)$ with $p = \frac{E_B - E_T}{2 \sum_j \alpha_j}$. That is, the leaving player k has a positive incentive to stay if and only if

$$\alpha_k \frac{(E_B - E_T)^2}{8(\sum_{j=1}^{N} \alpha_j)^2} < \delta_k. \tag{3}$$

That is, the mitigation burden borne by player k (left-hand side) based on the remaining coalition keeping the threshold must be smaller than its damages δ_k. That is, the "pivotal" player k has a higher incentive to stay in the grand coalition if (i) δ_k is large (ii) α_k is small (his mitigation potential is low) or (iii) α_k is very large (i.e., α_k in the denominator dominates, so that the coalition's total mitigation costs are sufficiently low). Intuitively, the non-linear effect of the marginal abatement costs α_k reflects the fact that on the one hand, for small values of α_k, the country has high mitigation costs and hence does little mitigation in the grand coalition, while for a very large value of α_k, on the other hand, keeping the threshold in the grand coalition is relatively easy if region k is a member and achieves keeping the threshold.

3 Implementation of thresholds in numerical coalition formation models

The previous section depicts by means of a simple model how the coalition formation process critically depends on the parameter values of the location and damage costs of a threshold. Here, we apply two empirically calibrated integrated assessment models (IAMs) to see how this strategic behavior plays out in two real world calibrated models, MICA and WITCH. Both models derive economic strategies with respect to climate change mitigation from an optimal growth framework. The models combine the two-stage game described above with an integrated climate economy model in the second stage.

The Model of International Climate Agreements (MICA, Lessmann et al., 2009, 2015) follows the same economic framework as RICE (Nordhaus and Yang 1996) but with different assumptions about mitigation costs and damage costs. It relies on stylized mitigation cost functions to model emissions reductions and neglects inertias in investing in mitigation technologies. In contrast, WITCH incorporates an explicit representation of mitigation options, particularly in the energy system (Bosetti

et al., 2006, Emmerling et al., 2016). More detailed model summaries are found in the Appendix B of the Supplementary Information available at the journal's website.

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):

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

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.

In both models, the following additional threshold-like function was added to $D(i,t)$, in accordance with the simple specification in equation (2). We use the cumulative distribution function of the normal distribution (known as the "error function", abbreviated erf) as a smooth (differentiable) approximation of step function thresholds (in the limit as $\sigma \to 0$). T_s is the location of the threshold[5] 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. Finally, d is the maximum damage from crossing the threshold, as a share of GDP, which –due to lack of further empirical evidence – is assumed to be symmetric across all regions. Taken together, this term can be written as $d * \text{erf}((T(t) - T_s)/\sigma) * \text{GDP}(i,t)$.

For the following runs we fixed $\sigma = 0.05$, which induces a continuous function that is very close to a step of magnitude d in damages (for $T_s = 2.5$ and $d = 0.04$, the damage at one standard deviation below, i.e., at $T(t) = 2.45$ is only $d = 0.0031$). The location of the threshold, T_s, and the maximum damages, d, were varied. For most of the runs, d was set such that 4% of GDP would be lost each period following the crossing of the threshold. While little is known about the economic impacts of crossing a tipping point, values used so far include the range of 5% to 10% of GDP (Cai et al., 2016), and values based on historical "catastrophic" GDP losses suggest extreme values of up to 20% (Barro and Jin, 2011). The final damage costs that enter the budget equation are:

$$D(i,t) = [\Omega(i, T(t)) + d * \text{erf}((T(t) - T_s)/\sigma)] * \text{GDP}(i,t)$$

In order to find the equilibrium in emission strategies in the second stage of the game, both models perform a fixed point iteration in emission strategies. Each iteration updates the emissions of non-signatories to maximize their individual welfare given the emissions of all other regions, and the emissions of the coalition members maximizes joint welfare to internalize all climate change externalities among coalition members. We found that threshold damages give rise to multiple equilibria in emission strategies in MICA. We have thus performed a systematic equilibrium selection process, which is described in detail in Appendix C. The results presented here are for the equilibrium in which the coalition attains their highest aggregate welfare.

[5] Equivalent to the threshold in terms of cumulative emissions E_T of the previous section, given that temperature increase and cumulative emissions have an almost linear relationship (Matthews et al., 2009).

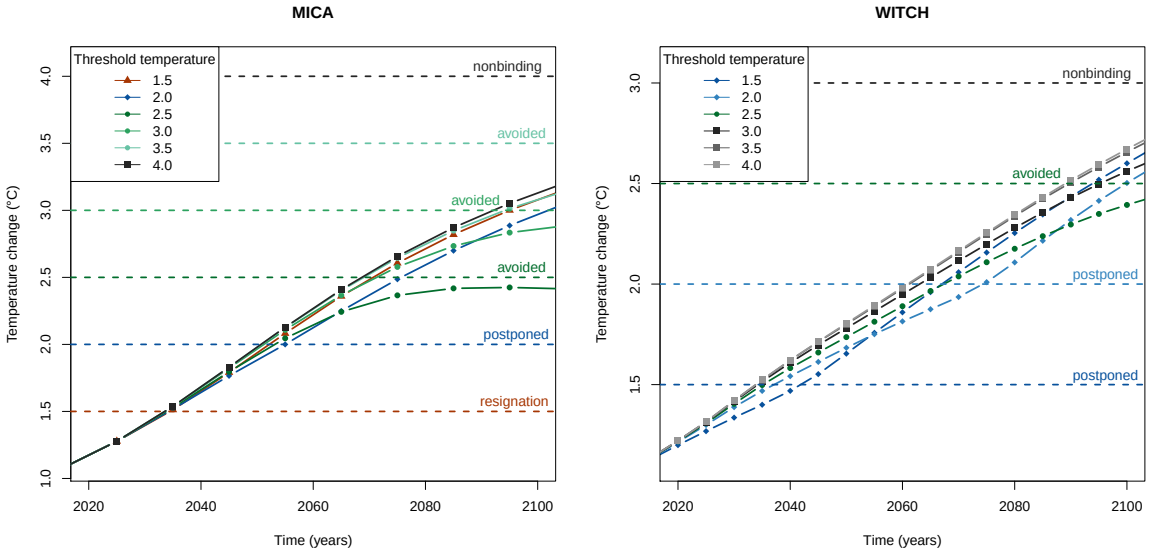

Figure 1: Temperature over time for different locations of the threshold T_s for MICA (left) and WITCH (right), $d = 0.04$ and $\sigma = 0.05$

4 Results

Our analytical results suggest that deterministic climate thresholds can enhance cooperation for certain parameter values of the model. Here, we use our numerical models to explore the role of thresholds on the second stage of the game, looking at how coalitions adjust their emissions and whether they keep temperatures below the threshold. Then we discuss the role of thresholds at the membership stage of the game.

4.1 Mitigation behaviors of coalitions

Figure 1 shows emissions from the two models for different threshold locations. We observe the four coalitions' emissions strategies identified in Section 3: (1) nonbinding, (2) avoidance, (3) postponement and (4) resignation. The scenarios in Figure 1 displays socially optimum behavior, i.e., the grand coalition strategies.

In MICA, the grand coalition keeps the thresholds for temperatures above or equal to $T_s = 2.5°C$. For $T_s = 2°C$ or lower, staying below the threshold is too costly and therefore either postponing or ignoring threshold strategies can be observed. For $1.5°C$ and lower temperature thresholds, which are bound to be crossed in the next decades, the effort required for postponing them is larger than the benefits. On the other extreme, for $T_s = 4°C$ and higher, the temperature increase in 2100 is just below or at the temperature that the grand coalition would keep without additional mitigation, thus the threshold is nonbinding. Therefore, the black curve represents the temperature profile the grand coalition of all regions would achieve in the absence of thresholds. In WITCH, we confirm this pattern, the only difference being that it is for thresholds lower than $1.5°C$ that the postponement behavior may be observed.

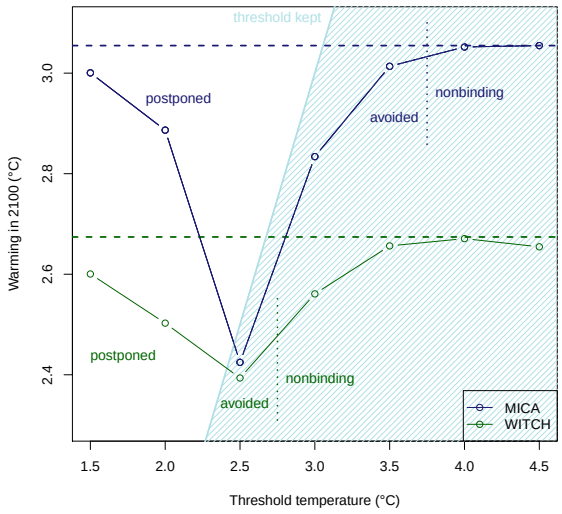

Figure 2: Temperature in 2100 reached by the grand coalition of all regions for different locations of the threshold, $d = 0.04$ and $\sigma = 0.05$

Looking at the overall results in both models in the equilibrium starting from the non-cooperative solution, the different regimes of behavior are summarized in Figure 2, which displays the temperature in 2100 for different locations of the threshold. When the location of the threshold T_s is low, keeping the threshold is too costly for the coalition and temperatures in 2100 thus exceed T_s. For higher threshold locations T_s, it pays for the coalition to postpone the time of exceeding the threshold (sometimes until the end of the time horizon), resulting in a lower temperature in 2100. When T_s falls on a temperature that is high enough, the coalition will keep the temperature below T_s in 2100. These results can be understood with the help of Section three: If keeping the threshold is too costly (high mitigation costs or low $\sum_j \alpha_j$ in comparison to the damage costs $\sum_j \delta_j$ it induces), emissions will increase to the level without the presence of a threshold. In the numerical models, this decision is spread over the entire time-horizon and can be taken for each time-period. Hence, we observe a much more nuanced postponement behavior of coalitions. This non-linear relation between the effects of thresholds on cooperation resembles an inverse U-shaped curve. Moreover, the nonlinear relationship seems to mirror similar results in environmental research in the amount of regime shifts on cooperation in the case of fisheries found in (Miller and Nkuiya, 2016) and the optimal ecosystem load of pollutants (Brozović and Schlenker 2011).

Mitigation behaviors upon defection

The grand coalition is indicative of the socially optimal behavior at climate thresholds. For the stability of climate agreements, the strategic reaction to defection by the remaining coalition is key. We consider the subcoalitions to the grand coalition to investigate these strategic responses to free-riding. Figure 3 shows the change in cumulative emissions by the (remaining) coalition members when the player denoted in the figure legend leaves the grand coalition. We focus on the defection of three key regions

Africa, China, and Russia in the first row as they exemplify types of behaviors (and incentives). The second row adds the response to when the regions making up the OECD leave the grand coalition. All subcoalitions are depicted in Figure S4 in the Appendix (available online in the Supplementary Information). The response of the remaining regions in the coalition may be to increase or decrease emissions to some extent, and this crucially depends on whether the regions remaining in the coalition have the ability of keeping the threshold. This in turn is a function of the stringency of the threshold and the amount of emissions the defecting region produces. At $T_s = 1.5$ and 2.0, where the threshold is not kept by the grand coalition, we see a slight increase in emissions by the remaining regions: this is the regular response to free-riding in models without threshold damages. Beginning at $T_s = 2.5$, however, the threshold is kept by the grand coalition in both the MICA and WITCH model. In MICA, coalition emissions skyrocket by up to 600 GtC for many coalitions upon defecting of a single region. Here, the remaining coalition abandons the previously avoided threshold. In particular, Africa and China are pivotal to keeping the threshold at $2.5°C$. Figure S5 in Appendix D shows that in MICA the defection of all regions but Japan, Russia and Rest of the World triggers a large increase in emissions. The mitigation potential of these pivotal regions is large, rendering the costs of keeping the threshold too high without their participation. MICA exhibits a simple structure of mitigation which abstracts from investment dynamics in emission-free capital. This allows for large responses to the threshold when a region leaves the grand coalition.

Contrary, when Russia free-rides in Figure 3, the remaining coalition goes on to still avoid the threshold by reducing their emissions. As noted in the Introduction, this behavior is contrary to the regular response to free-riding in models with continuous damages were assumed. Moderate changes in cumulative emissions indicate that the postponement behavior is optimal for some coalitions and emissions are only marginally changed when the coalition becomes smaller (see for example Africa for $T_s = 3°C$).

Although in WITCH the changes in emissions are less pronounced (cf. Figure 3 right), qualitatively the same patterns emerge. The changes overall are lower in terms of emission differences, since in WITCH modeled investment dynamics and hence inertia in the energy system make extreme changes in mitigation very costly. In addition, due to its numerical complexity WITCH is solved with a time horizon until 2100 while MICA has a time horizon until 2195. For low threshold locations, the remaining coalition raises its emission level, peaking at $T_s = 2.5$, which is avoided by the grand coalition, but exceeded if China or Sub-Saharan Africa leave the coalition. These players are thus pivotal. At $T_s = 3.0$ and $T_s = 3.5$ we see the same strategic behavior as in the MICA model: in contrast to the regular response to free-riding, the remaining coalition emits less after defection of a single region, except for Sub-Saharan Africa. With higher threshold locations, this effect is then again replaced by the regular free-riding response, as the threshold temperature becomes non-binding and thus loses importance for the emissions behavior.

4.2 Stability results

The type of strategic behaviors in emissions just described affects the stability of coalitions, in particular around critical threshold temperatures. Figure 4 shows the value of the stability function for the three important regions Africa, China, and Russia in the first row and the regions making up the OECD in

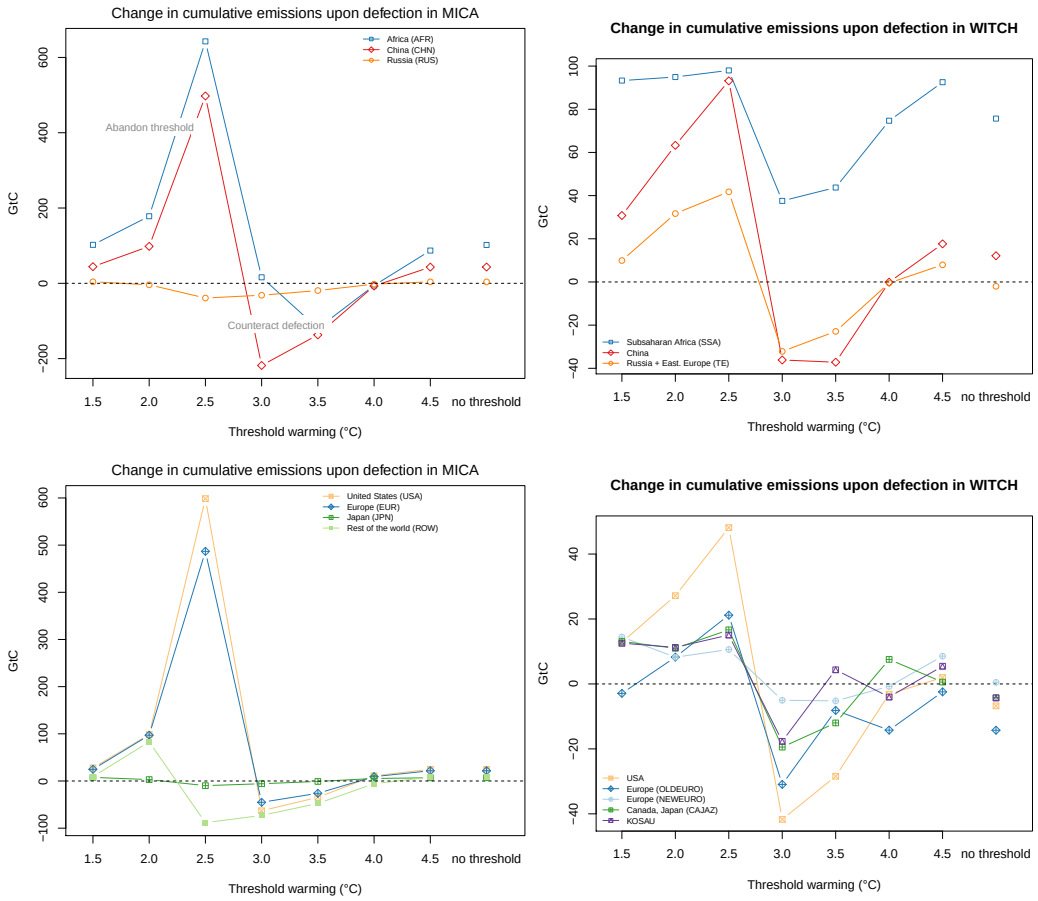

Figure 3: Changes in cumulative emissions (over the 21st century) for MICA and WITCH for different temperature thresholds (selected regions)

the second row. For both models, the value of the stability function is generally negative, similarly to the case in the absence of threshold damages. This is expected, since being the sole free-riding region against what remains of the grand coalition is highly beneficial for the free-rider, and the incentive to free-ride is therefore large. At the critical threshold temperature of $T_s = 2.5$, we see a dramatic change in the MICA model. The threshold at $T_s = 2.5$ is avoided by the grand coalition, but (as we know from Figure 3) not anymore upon defection by Africa. The prospect of exceeding the threshold creates the incentive for China and Russia to rather remain in the grand coalition. Figure S4 in the Appendix shows that the incentive to remain is also positive for the regions India, Other-Asian, USA, Europe and Latin-America at $T_s = 2.5$.

This positive effect on the willingness to cooperate is quickly lost for higher threshold levels. Worse, when both the grand coalition and the remaining subcoalition avoid the threshold, free-riding becomes even more profitable than without threshold damages, i.e. the stability value falls below this benchmark case. By defecting, the free-rider lowers its individual mitigation costs while the damage level remains virtually unchanged due to the increased effort of the remaining coalition (cf. Figure 3). At $T_s = 3.0$, all regions have a negative incentive to remain a member to the grand coalition (Figure S4). These strategic effects of anticipating the remaining coalition to abandon or maintain the threshold are present in the analytical model and are confirmed in the numerical analysis. Moderate changes to the mitigation 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.

China is pivotal to keeping the threshold at $2.5°C$ but still has an incentive to leave the grand coalition. In line with the theoretical model its mitigation costs are too high when the threshold is kept inside the coalition compared to its individual avoided damages. The same holds for the region of North African and Middle Eastern Countries (MEA): while being pivotal to keeping the threshold at $2.5°C$ in MICA, its incentive to remain a member of the grand coalition is negative.

We saw above that in WITCH the changes in emissions were less pronounced. This is mirrored by a smaller impact on the stability function, which changes sign in only a single case: For Sub-Saharan Africa at $T_s = 3.5$, the prospect of defecting while the remaining coalition members make up for its increase in emissions is too tempting to remain in the grand coalition. Therefore, the values of the stability function are negative for most regions and scenarios, see Figure 4 (right) and Figure S4 in the Appendix for all regions. Only for regions with relatively high damages relatively low mitigation costs (Sub-Saharan Africa, India, and South Asia), the stability function in the grand coalition shows a positive value.

4.3 Pivotal regions

Due to the heterogeneity of players in both numerical models, the presence of thresholds does not induce stability of the grand coalition in any scenario. In MICA and WITCH, a positive incentive to sign the agreement is found for some regions only. The regions that have an incentive to stay inside the grand coalition have the following characteristics: first and foremost, the mitigation 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, so that the threshold is abandoned

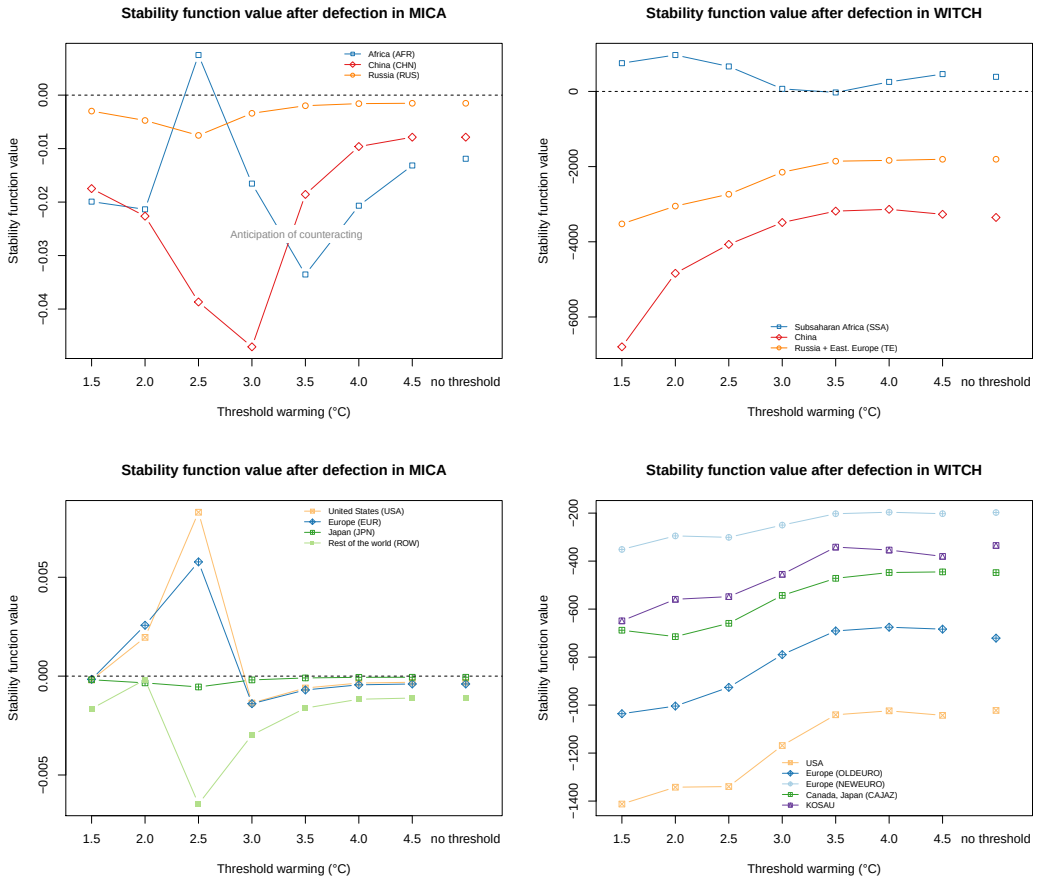

Figure 4: Stability functions for MICA and WITCH for different temperature thresholds (selected regions)

by the remaining coalition. These regions are pivotal in the sense that their membership is necessary to keep the threshold. At a threshold temperature of $2.5°C$, we find that all regions but Japan, Russia and Rest of the World are pivotal in MICA. In WITCH, Sub-Saharan Africa (SSA), India, and South Asia are non-pivotal players while all other players are necessary to be in the coalition to keep the threshold (see Figure S4 in the Appendix). Secondly, increased mitigation costs need to be valued against the benefits of keeping the threshold for an individual region. If damages are not sufficiently high and individual emissions reductions are too costly in comparison, leaving the coalition can become attractive, even if the threshold is crossed upon leaving (China and MEA in the MICA model for a threshold temperature of $2.5°C$ for example). The endogenous interplay between mitigation and damage costs therefore determines stability in a complex manner. For the threshold to set an incentive to stay for a region, these regions or countries need to be pivotal to keep the threshold, but in addition their individual benefits from keeping the threshold need to be high enough.

4.4 Alternative stability concepts

The equilibrium concept analyzed so far in this study is internal and external stability of international environmental agreements. Internal/external stability is myopic with respect to the membership decision: when a player considers defection from any given coalition, this player will assume continued cooperation of the remaining coalition members ignoring any subsequent (or simultaneous) defections by other coalition members. The pessimistic results from this approach can (partially) be traced back to this myopia of the defecting player (Finus, 2003). Concepts of farsighted coalition stability address this concern by considering chains of subsequent defections (Aart de Zeeuw, 2008). This raises the cost of defection, as the point of comparison for defecting from an N player coalition is not cooperation of $(N-1)$ players but potentially a much smaller coalition. Taken to the extreme, the point of comparison could become the non-cooperative equilibrium. In this case, it would be enough to sustain cooperation in a coalition S, if participation is *individually profitable* for all members, i.e., $\pi_i(S) > \pi_i(S^{\mathrm{NC}})$ where S^{NC} is the non-cooperative equilibrium where all coalitions are singletons. Thus, profitability marks the polar case to the myopic internal/external stability concept on the spectrum ranging from the optimistic expectation of continued cooperation of all remaining members to the pessimistic assumption of a complete break-down of cooperation. Additionally, individual profitability is also a necessary condition for stability in the sense of the γ-core, which identifies coalitions where no subcoalition (blocking coalition) does better for all its members – as a non-profitable player constitutes a singleton blocking coalition. Chander and Tulkens (1997) analyze core stability for economies with externalities with transferable utility; for non-transferable utility (as in this study), blocking coalitions need to do better member-by-member as defined in Myerson (1991, Ch. 9.8).

To investigate how this equilibrium concept affects coalition stability under climate thresholds, we compute individual profitability for the grand coalition.[6] Figure S6 in the Appendix summarizes individual profitability for a range of temperature thresholds locations. Overall, profitable does not change drastically over different thresholds and avoiding the threshold is in many cases individually profitable. We find that in WITCH, irrespective of the threshold, there are only four regions for which

[6] For rigorous tests of farsighted or γ-core stability a full set of all possible coalitions is needed but is not available as the additional computational effort puts this beyond the scope of this study. A discussion of testing core stability in non-transferable utility models is found in Kornek et al. (2014).

the grand coalition is profitable (India, Africa, South Asia, and South East Asia). Hence the grand coalition with socially optimal strategies cannot be stable in the sense of the γ-core, as the remaining majority of regions do better in the non-cooperative equilibrium and could block the grand coalition. Still, profitability paints a slightly more optimistic picture regarding the incentive to cooperate, as East Asia – in contrast to the other three profitable regions – does not have a positive value of the stability function (see Figure S4 in the Appendix).

For MICA, profitability is almost always given. Profitability peaks at the $T_S = 2.5$ threshold from whereon the grand coalition keeps the temperature below the threshold T_S and cooperation thus provides a great benefit. For lower values of T_S, profitability is much lower (and for $T_S = 2.0$, we have the only case where profitability is negative in the case of Africa). The prospect for cooperation is thus much improved in this model when departing from the myopic internal/external stability perspective, where – except at the critical $T_S = 2.5$ – the standard stability function was negative for the majority of regions (see Figure S4). Therefore, based on these results stability of the grand coalition with socially optimal strategies according to the γ-core can only be ruled out at $T_S = 2.0$, while for higher thresholds individual profitability is given.

The greater scope for cooperation following from farsightedness has been shown in theory (Chander, 2007). This exercise shows that farsightedness translates to a difference in the prospect for cooperation in numerical models that is substantial (in MICA more pronounced than in WITCH). A rigorous implementation of farsighted stability would plausibly fall in-between the polar cases of internal/external stability and profitability, such that the dynamics of pivotal players at a critical threshold, as discussed above, may make the difference whether a coalition is indeed stable.

4.5 Stability with transfers for different threshold locations and damages

In the MICA model, at $T_s = 2.5°C$ and threshold damages of 4 per cent ($d = 0.04$), Figure S4 in the Appendix shows that six 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 remaining five regions that lose from cooperating, stability of the grand coalition could be achieved. We test if there exist transfers that once implemented realize a positive incentive to sign for all regions using the method from Kornek et al. (2014). Table 1 shows the combinations of location of the threshold and maximum damage costs where there exists a transfer scheme within the grand coalition such that a positive incentive to sign for all regions is attained. For nine out of these 45 scenarios the gains of cooperation that accrued in some regions were enough to compensate all regions that lose from cooperation.

In columns with low threshold damages of $d \leq 0.025$ and hence low gains from cooperation, no transfer scheme was sufficient to compensate all losers. Only at higher threshold damages do we find transfer schemes that make the grand coalition stable. These stable coalitions (with transfers) are, however, restricted to a narrow band of threshold locations starting at $T_S = 2.5$ for $d \geq 0.03$ and shifting towards lower threshold locations with increasing maximum damage costs. The intuition for this narrow band is this: Threshold temperatures T_S below the band necessitate extremely ambitious emissions reductions such that even optimal strategies of the grand coalition will exceed this threshold. Emissions strategies then revert back to the traditional free-riding behavior as with continuous damages. A strong incentive to leave results in all regions, thus diminishing the scope for transfers to enhance cooperation.

	Maximum damage costs d (percent)								
Threshold location T_S	0.02	0.025	0.03	0.035	0.04	0.045	0.05	0.055	0.06
2.00	0	0	0	0	0	0	0	1	0
2.25	0	0	0	0	0	1	1	1	1
2.50	0	0	1	1	1	1	0	0	0
2.75	0	0	0	0	0	0	0	0	0
3.00	0	0	0	0	0	0	0	0	0

Table 1: 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 ($\sigma = 0.05$)

On the other side with threshold temperatures above the band, keeping the threshold is feasible for more subcoalitions after a single region defects from the grand coalition. When subcoalitions abate ambitiously, the additional gains of cooperation in the grand coalition are low, such that there are too few regions with an incentive to sign the agreement. Therefore, only a narrow band of threshold temperatures induces sufficiently many regions to have a positive incentive to sign the agreement, making compensation of the other regions possible. The band shifts towards lower threshold locations for higher maximum damage costs: for lower and thus more ambitious threshold temperatures the grand coalition will only keep the threshold for higher maximum damages. This moves the "band" where the grand coalition keeps the threshold upwards in Table 1 for higher values of the parameter d.

5 Extension to uncertain thresholds

As discussed in section 1, uncertainty may have a crucial effect on the stability of coalitions. In this section, we exemplary test how introducing uncertainty may affect our findings. The preceding section 4.5 showed that the location T_S of the threshold temperature is critical for the threshold to make a difference in the participation decision. While an uncertain threshold location in the model would be interesting to evaluate its impact for the stability of coalitions, in the numerical models used here, its implementation is virtually impossible. Hence, here we focus on the critical threshold locations established in this study ($T_S = 2.5$ and $T_S = 3.0$ for the MICA and WITCH models, respectively) and explore uncertainty by considering two polar cases: (a) full impact of threshold damages at the critical temperature (i.e., $d = 0.04$ as throughout this study) and (b) no damage is triggered at the threshold (i.e., $d = 0.0$). We assume uncertainty about threshold damages in the sense of equal probability p for the two cases and assess the impact of this uncertainty by considering *decision making under uncertainty* in the participation stage.[7] Uncertainty is resolved before investment and mitigation decisions are made in stage two. That is, the participation decision in stage one is taken based on the expected utilities that a regions would have inside or outside the coalition, where the expectation is

[7] This uncertainty about threshold damages is conceptually equivalent to the following uncertainty about the threshold location. A threshold with damage $d = 0.04$ materializes at the temperature $T_S = 2.5$ (and $T_S = 3.0$ respectively) or at an infinitely large temperature, with 50% probability each. See Barrett (2013) for a discussion about the different implications of damage vs. threshold uncertainty.

Scenario 1	Threshold	$d = 0.04$
Scenario 2	No threshold	$d = 0.00$
Scenario 3	Expected utility	p(Scenario 1) + $(1-p)$(Scenario 2)

Table 2: Scenario overview for the uncertainty analysis

taken over the alternative realizations of stage two for different values of d. Table 2 summarizes the scenarios.

We find that when decisions are made based on expected utility, i.e., the average welfare in scenarios 1 and 2 weighted by their probabilities, stability is impeded in MICA and WITCH. Figure 5 illustrates this by way of example for the grand coalition. We show the value of the stability function $\varphi_i(N)$ for all members i of the grand coalition N in terms of expected utility, accompanied by the underlying polar cases with full or no threshold damage. Uncertainty reduces the prospect for cooperation, first and foremost, because the possibility of a world without a climate damage threshold reduces the expected stability value for every region. Put differently, free-riding on 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.

A comparison of the Threshold and No Threshold scenarios for MICA shows how the existence of threshold damages flips the incentive for Africa (AFR), Latin America (LAM), India (IND), other Asian countries (OAS), USA and Europe (EUR) in favor of participating in the coalition. This effect of threshold damages is reduced by uncertainty but only for AFR and LAM does it revert the incentive back to non-participation. For a total of four regions (including India, USA and Europe), the benefits from avoiding an (uncertain) threshold still outweigh the gains from free-riding in the absence of a threshold. Hence, the presence of the uncertain threshold may still increase the scope for cooperation. The effect of uncertainty has the same direction in WITCH but is more nuanced, i.e., the sign of the stability function is not affected for any region.

We assumed equal probability for the two polar scenarios. Other choices for the probability p would shift the expected utility towards either of the polar cases. Further investigations and more levels of uncertainty, in particular about the threshold location and for higher potential impacts, are necessary. Our model nevertheless illustrates the difficulties that arise once thresholds are uncertain in terms of location or magnitude.

6 Conclusion

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 since countries find themselves in a classical "Prisoner's Dilemma". Recent literature has shown how free-riding incentives are overcome when thresholds in the damage costs are considered in the analysis. Here, we find that this result is very sensitive to the characteristics of the threshold considered.

The numerical analyses with the models MICA and WITCH show that the socially optimal emissions – when all regions cooperate – keep temperatures below a threshold of moderate warming (ap-

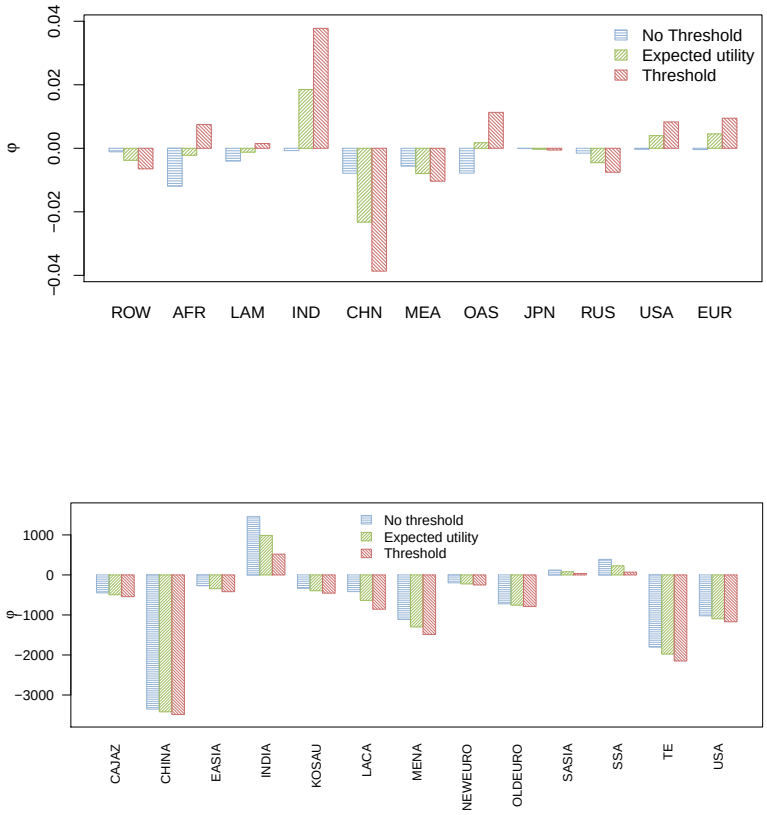

Figure 5: Value of the stability function for four different scenarios: Without the presence of a threshold (blue); with a presence of a threshold (red); Expected Utility (green); MICA in the upper figure and WITCH in the lower figure with $T_S = 2.5°C, d = 0.04, \sigma = 0.05$

proximately at 2.5 to 3.0°C) and of sufficiently large damage costs (several percentage points of GDP). Otherwise, mitigation costs are too high compared to the damage costs so that keeping the threshold is not Pareto optimal. The resulting optimal temperature thus follows a U-shaped relationship with the threshold. This non-monotonicity could also explain why the discussion about catastrophic climate impacts has led to different arguments for the implications for mitigation. If one region defects from the grand coalition of all regions, it may be optimal for the remaining signatories to either keep the threshold for the entire time horizon, stay below the threshold temperature only temporarily or increase emissions 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 subcoalition finds it optimal to keep the threshold, emissions actually decrease 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 while the costs of emission reductions decrease significantly. Hence, cooperation is impeded in this case. If, on the other hand, the subcoalition increases emissions 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 mitigation costs upon leaving is not too high compared to the increase in damage costs, pivotal regions may find it optimal to sign the agreement.

Our results show that in particular the location of climate thresholds is critical to shape incentives. Threshold locations at around 2.5°C enhance cooperation if the potential damage of crossing the threshold is in the order of a few percentage points of GDP. We find that if compensation between regions is possible, the grand coalition of all regions can be stable for this combination of threshold location and damage size. However, diverging from location or impact level of the threshold can reverse this conclusion: while a tipping point at a very low temperature threshold can lead to a mere postponement of passing the threshold or total ignorance/resignation, thresholds at (much) higher temperatures become non-binding and thus don't change the incentive structures and mitigation outcomes. Assessing the effect of threshold damages on cooperation therefore hinges on research shedding light on the location of the threshold and potential damages associated with it. Further research is necessary to investigate the numerical characteristics of the various potential and uncertain climate thresholds. In particular, analyzing the case of uncertain threshold locations for coalition formation would be highly relevant also for numerical applications, see, e.g., Lemoine and Traeger (2016). Finally, we also discuss how uncertainty affects the analysis of this paper.

7 References

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

Barrett, S., 2003. Environment and statecraft: The strategy of environmental treaty-making: The strategy of environmental treaty-making. Oxford University Press, Oxford.

Barrett, S., 1994. Self-Enforcing International Environmental Agreements. Oxford Economic Papers 46, 878–94.

Barrett, S., Dannenberg, A., 2012. Climate negotiations under scientific uncertainty. Proceedings of the National Academy of Sciences 109, 17372–17376.

Barro, R.J., Jin, T., 2011. On the Size Distribution of Macroeconomic Disasters. Econometrica 79, 1567–1589.

Benchekroun, H., Van Long, N., 2012. Collaborative environmental management: a review of the literature. International Game Theory Review 14, 1240002.

Bosetti, V., Carraro, C., Galeotti, M., Massetti, E., Tavoni, M., 2006. WITCH A World Induced Technical Change Hybrid Model. The Energy Journal 27, 13–37.

Brozović, Nicholas, and Wolfram Schlenker, 2011. Optimal Management of an Ecosystem with an Unknown Threshold. Ecological Economics 70 (4), 627–40.

Cai, Y., Lenton, T.M., Lontzek, T.S., 2016. Risk of multiple interacting tipping points should encourage rapid CO2 emission reduction. Nature Clim. Change 6, 520–525.

Chander, P. 2007. The Gamma-Core and Coalition Formation. International Journal of Game Theory 35 (4), 539–56.

Chander, P., 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., Tulkens, H., 1997. The Core of an Economy with Multilateral Environmental Externalities. International Journal of Game Theory 26 (3) , 379–401.

Diekert, Florian K., 2017.Threatening Thresholds? The Effect of Disastrous Regime Shifts on the Non-Cooperative Use of Environmental Goods and Services. Journal of Public Economics 147, 30–49.

D'Aspremont, C., Gabszewicz, J.-J., 1986. On the stability of collusion, in: New Developments in the Analysis of Market Structures. Macmillan, New York, Ch. On the stability of collusion, pp. 243–264.

Emmerling, J., Drouet, L., Reis, L.A., Bevione, M., Berger, L., Bosetti, V., Carrara, S., Cian, E.D., D'Aertrycke, G.D.M., Longden, T., Malpede, M., Marangoni, G., Sferra, F., Tavoni, M., Witajewski-Baltvilks, J., Havlik, P., 2016. The WITCH 2016 Model - Documentation and Implementation of the Shared Socioeconomic Pathways, FEEM Working Paper No. 2016.42, Fondazione Eni Enrico Mattei.

Finus, M., 2003. New Developments in Coalition Theory: An Application to the Case of Global Pollution. In: Marsiliani, L., M. Rauscher and C. Withagen (eds.), Environmental Policy in an International Perspective, Kluwer Academic Publishers, Dordrecht, Holland, pp. 19-49.

Finus, M., 2008. 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.

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

Hoel, M., 1992. International environment conventions: the case of uniform reductions of emissions. Environmental and Resource Economics 2, 141–159.

Iris, D., Tavoni, A., 2016. Tipping Points and Loss Aversion in International Environmental Agreements, FEEM Working Paper No. 25.2016, Fondazione Eni Enrico Mattei.

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

Kornek, U., Steckel, J.C., Lessmann, K., Edenhofer, E. 2017. The climate rent curse: new challenges for burden sharing. Int Environ Agreements 17 (6), 855–882.

Kornek, U., Lessmann, K., Tulkens, H., 2014. Transferable and non transferable utility implementation of coalitional stability in integrated assessment models. CORE working paper 2014/35.

Kriegler, E., Hall, J.W., Held, H., Dawson, R., Schellnhuber, H.J., 2009. Imprecise probability assessment of tipping points in the climate system. PNAS 106, 5041–5046.

Lemoine, Derek, and Christian P. Traeger, 2016. Ambiguous Tipping Points. Journal of Economic Behavior & Organization, 132, 5–18.

Lenton, T.M., Held, H., Kriegler, E., Hall, J.W., Lucht, W., Rahmstorf, S., Schellnhuber, H.J., 2008. Tipping elements in the Earth's climate system. Proceedings of the National Academy of Sciences 105 (6), 1786-1793.

Lessmann, K., Kornek, U., Bosetti, V., Dellink, R., Emmerling, J., Eyckmans, J., Nagashima, M., Weikard, H.-P., Yang, Z., 2015. The Stability and Effectiveness of Climate Coalitions. Environmental and Resource Economics 62, 811–836.

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

Lontzek, T.S., Cai, Y., Judd, K.L., Lenton, T.M., 2015. Stochastic integrated assessment of climate tipping points indicates the need for strict climate policy. Nature Clim. Change 5, 441–444.

Mäler, Karl-Göran, Anastasios Xepapadeas, and Aart de Zeeuw, 2003, The Economics of Shallow Lakes. Environmental and Resource Economics 26 (4), 603–624.

Matthews, H. Damon, Nathan P. Gillett, Peter A. Stott, and Kirsten Zickfeld, 2009. The Proportionality of Global Warming to Cumulative Carbon Emissions. Nature 459, 829–32.

Miller, S., Nkuiya, B., 2016. Coalition formation in fisheries with potential regime shift. Journal of Environmental Economics and Management 79, 189–207.

Muradian, Roldan, 2001. Ecological Thresholds: A Survey. Ecological Economics 38 (1), 7–24.

Myerson, R.B., 1991. Game theory: analysis of conflict. Harvard University Press.

Nagashima, Miyuki, Rob Dellink, Ekko van Ierland, and Hans-Peter Weikard, 2009. Stability of International Climate Coalitions - A Comparison of Transfer Schemes. Ecological Economics 68 (5): 1476–87.

Nordhaus, W.D., 1994. Managing the global commons: the economics of climate change. MIT press Cambridge, MA.

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

Polasky, Stephen, Aart de Zeeuw, and Florian Wagener. 2011. Optimal Management with Potential Regime Shifts. Journal of Environmental Economics and Management 62 (2): 229–40.

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

Rogelj, Joeri, Michel den Elzen, Niklas Höhne, Taryn Fransen, Hanna Fekete, Harald Winkler, Roberto Schaeffer, Fu Sha, Keywan Riahi, und Malte Meinshausen, 2016. Paris Agreement climate proposals need a boost to keep warming well below 2°C. Nature 534, 631–39.

Sakamoto, Hiroaki, 2014. Dynamic Resource Management under the Risk of Regime Shifts. Journal of Environmental Economics and Management 68 (1), 1–19.

Schmidt, Robert C., 2017. Dynamic Cooperation with Tipping Points in the Climate System. Oxford Economic Papers 69 (2), 388–409.

Tsur, Yacov, and Amos Zemel, 2016. Policy Tradeoffs under Risk of Abrupt Climate Change. Journal of Economic Behavior & Organization, 132, 46–55.

Weikard, H.-P., 2009. Cartel Stability under an Optimal Sharing Rule. The Manchester School 77, 575–593.

Zeeuw, Aart de, 2008. Dynamic Effects on the Stability of International Environmental Agreements. Journal of Environmental Economics and Management 55 (2), 163–74.

Online Appendix to the article Emmerling J., Kornek U., Bosetti V., and K. Lessmann, Climate thresholds and heterogeneous regions: implications for coalition formation. The Review of International Organizations, DOI: www.doi.org/10.1007/s11558-019-09370-0

A The Location of the threshold

The first empirical question hinges on the loci of thresholds in the climate system. The expert elicitations by Kriegler et al. (2009) provide a probabilistic assessment of the locations of important climate tipping points. These tipping points, also used in Cai et al. (2016), include: the reorganization of the Atlantic Meridional Overturning Circulation (AMOC); the melting of the Greenland ice sheet (GIS); the disintegration of the West Antarctic ice sheet (WAIS); the die-back of the Amazon rain forest (AMAZ); and the shift to a more persistent El Niño Southern Oscillation regime (ENSO).

Based on the hazard rates given in the original paper and following the method of Rinne (2014) one can derive the probability distribution associated to each tipping element. Figure S1 shows the distribution of those tipping elements for different degrees of global mean temperature increase over the average between 1980 and 1999. The median of the assessed distributions falls in the range between 2 to 4 degrees of warming, although part of the distribution falls below and above this range. For this study, we therefore consider the range 1.5-4.5 degrees of warming, which seems compatible with a range of reasonable thresholds in the climate system such as these five elements.

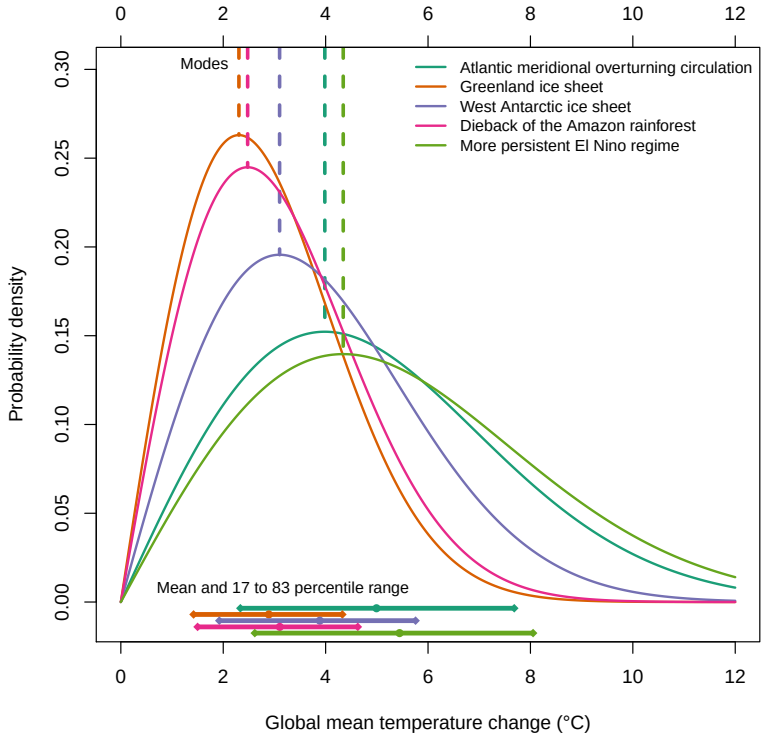

Figure S1: Different tipping points and their probabilistic temperature threshold

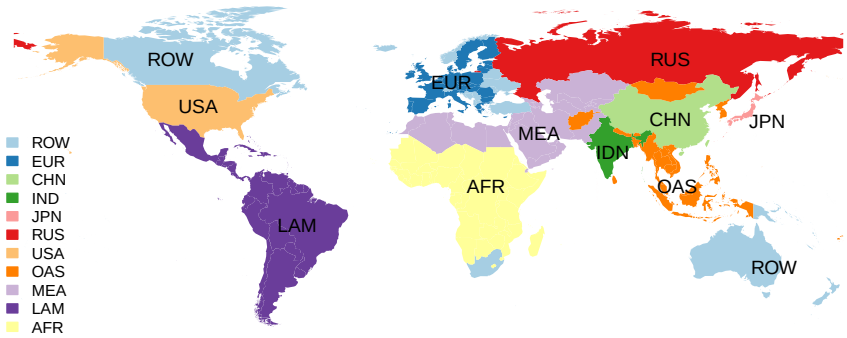

Figure S2: Regions represented as players in the MICA model

B Model descriptions

MICA (Model of International Climate Agreements) is a climate-economy model build to explore the effectiveness and the design of climate agreements. The model distinguishes eleven world regions, each described as a Ramsey-type optimal growth economy. It is similar to the seminal RICE model (Nordhaus and Yang 1996) but in contrast draws information on mitigation costs from an emulation of the REMIND-R model (Luderer et al. 2013) and is flexible to use climate change damage modeling from a variety of studies, this paper uses the damage specification based on Fankhauser (1995). A modified Negishi algorithm allows including intertemporal trade when solving the model under assumptions of full or partial cooperation or non-cooperative behavior regarding the climate externality. A full description of the model equations is available as an appendix to Kornek et al. (2017).

WITCH (World Induced Technical Change Hybrid) is an integrated assessment model designed to assess climate change mitigation and adaptation policies. It is developed and maintained at the the RFF-CMCC European Institute on Economics and the Environment (EIEE). It is a global integrated assessment model with two main distinguishing features: a regional game-theoretic setup, and an endogenous treatment of technological innovation for energy conservation and decarbonization. A top-down inter-temporal Ramsey-type optimal growth model is hard linked with a representation of the energy sector described in a bottom-up fashion, hence the hybrid denomination. The regional and intertemporal dimensions of the model make it possible to differentiate and assess the optimal response to several climate and energy policies across regions and over time. The non-cooperative nature of international relationships is explicitly accounted for via an iterative algorithm which yields the open-loop Nash equilibrium between the simultaneous activity of a set of representative regions. Regional strategic actions interrelate through GHG emissions, dependence on exhaustible natural resources, trade of fossil fuels and carbon permits, and technological R&D spillovers. R&D investments are directed towards either energy efficiency improvements or development of carbon-free breakthrough

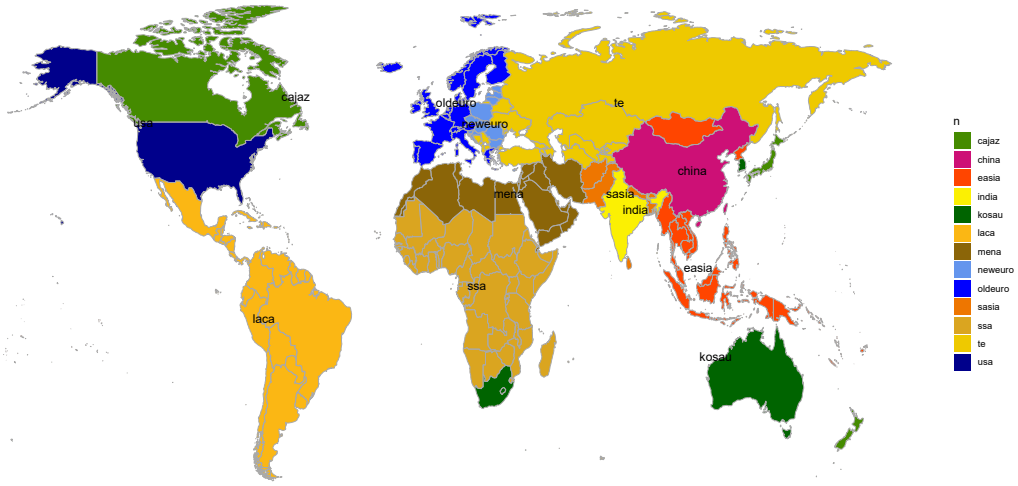

Figure S3: Regions represented as players in the WITCH model

technologies. Such innovation accumulates over time and spills across countries in the form of knowledge stocks and flows.

The competition for land use between agriculture, forestry, and bio-energy, which are the main land-based production sectors, is described through a soft link with a land use and forestry model (GLOBIOM, Global Biosphere Management Model. A climate model (MAGICC) is used to compute climate variables from GHG emission levels and an air pollution model (FASST) is linked to compute air pollutant concentrations. While for this exercise WITCH is used for cost-effective mitigation analysis, the model supports climate feedback on the economy to determine the optimal adaptation strategy, accounting for both proactive and reactive adaptation expenditures.

WITCH represents the world in a set of a varying number of macro regions – for the present study, the version with 13 representative native regions has been used; for each, it generates the optimal mitigation strategy for the long-term (from 2005 to 2100) as a response to external constraints on emissions. A model description is available in (Bosetti et al., 2006), and (Emmerling et al., 2016), and a full documentation can be found at `http://doc.witchmodel.org`.

C Equilibrium selection

It is known that we cannot rule out multiple equilibria when our models include non-convexities. This acknowledgment is often followed by the assertion that in standard applications of the model, multiple equilibria were not observed (e.g., Nordhaus and Yang 1996, Eyckmans and Tulkens 2003). However, Lempert et al. (2006) introduced abrupt changes in the climate dynamics into the DICE model and found multiple equilibria. This is no surprise as abrupt, threshold-like behavior implies strong non-convexities.

In MICA, when introducing threshold damages for this works, we found indeed different solutions by starting the solution algorithm with different initial values. At times the initial value determined

whether the threshold temperature limit was exceeded or kept.

For the purpose of this study, we chose to focus on the equilibrium solution with the highest social welfare. As the decision to exceed or keep the threshold temperature rests with the coalition, the equilibrium with the highest aggregate utilitarian welfare of all coalition members is selected. The remainder of this section provides the details of our equilibrium selection.

The numerical model implemented in GAMS/CONOPT (Drud (1996) searches for local solutions based on first-order conditions of optimality, i.e., based on marginal information. But marginal damage according to our threshold damage specification will be very similar left and right of the threshold. Our approach is therefore basically to compute and compare solutions where we enforce that the threshold temperature is kept or exceeded by an additional constraint. There is one caveat: as our model is dynamic, we need to vary the point in time when the threshold is exceeded.

Hence apply the following solution procedure for each coalition

1) We add the following two-part constraints to the model. The first part enforces that temperature stays below the threshold up to the threshold crossing time period χ. The second part forces temperature to rise beyond the threshold.

$$T(t) \leq T_S \qquad \forall t \leq \chi$$
$$T(t) > T_S \qquad \forall t > \chi$$

1) The model is run for all possible crossing time periods, including the case of keeping the threshold temperature over the full time horizon. The solutions--if feasible--of the thus constrained model are candidate solutions for the model without the additional constraints.

2) All candidate solutions are tested for the PANE property. This is done by running the optimization again for all regions and the coalition for the unconstrained model, keeping the decision variables of all other regions fixed. When all strategies remain unchanged under this re-optimization, the definition of the PANE is fulfilled.

3) Among all PANE solutions, the equilibrium solution with the highest welfare of coalition members (equally weighted aggregate) is selected.

In WITCH, on the other hand, we explored the possibility of different equilibria by running the model starting from the optimal GC or BAU equilibrium as starting value. We found for all runs that yet only one equilibrium was found in the different runs, yielding the global maximum in terms of welfare. This uniqueness has been previously found within this IAM, which covers a much higher degree of complexity and inertia resulting in less extreme possible solutions of the model.

D Stability functions for all model regions

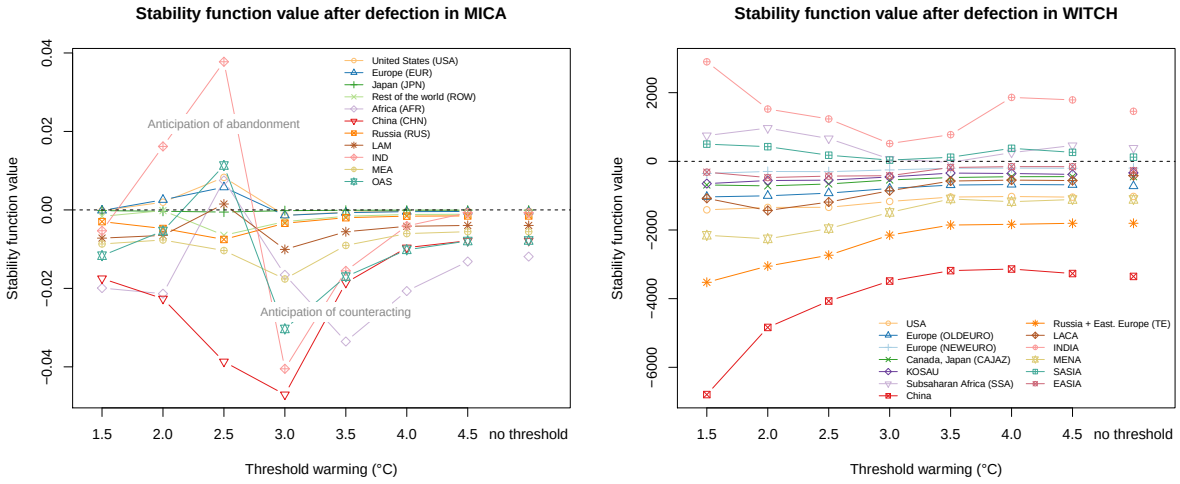

Figure S4: Stability functions for MICA and WITCH for different temperature thresholds

E Changes in cumulative emissions upon defection for all model regions

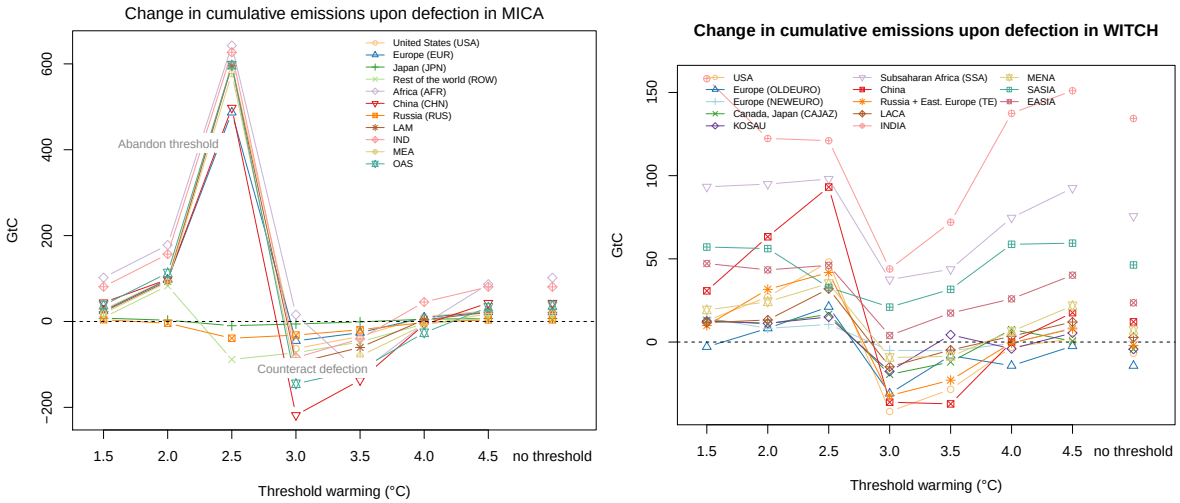

Figure S5: Changes in cumulative emissions for MICA and WITCH for different temperature thresholds

F Profitability for all model regions

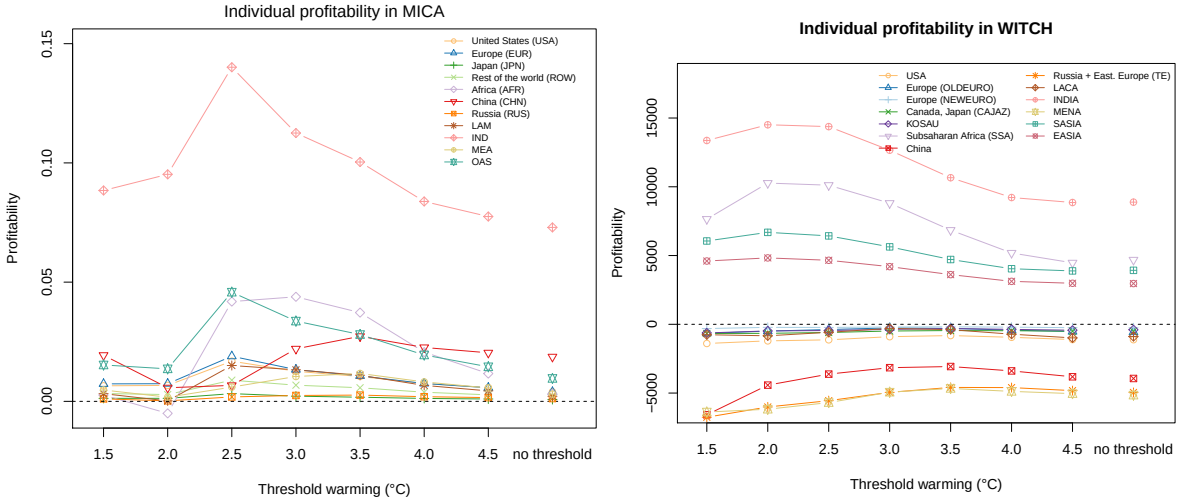

Figure S6: Profitability of the grand coalition for individual regions for MICA and WITCH at different temperature thresholds

G References

Bosetti, V., Carraro, C., Galeotti, M., Massetti, E., Tavoni, M., 2006. WITCH A World Induced Technical Change Hybrid Model. The Energy Journal 27, 13–37.

Cai, Y., Lenton, T.M., Lontzek, T.S., 2016. Risk of multiple interacting tipping points should encourage rapid CO2 emission reduction. Nature Climate Change 6, 520–525.

Drud, A.S. 1996. CONOPT: A System for Large-Scale Nonlinear Optimization, Reference Manual for CONOPT Subroutine Library. ARKI Consulting and Development A/S, Bagsvaerd, Denmark (1996).

Emmerling, J., Drouet, L., Reis, L.A., Bevione, M., Berger, L., Bosetti, V., Carrara, S., Cian, E.D., D'Aertrycke, G.D.M., Longden, T., Malpede, M., Marangoni, G., Sferra, F., Tavoni, M., Witajewski-Baltvilks, J., Havlik, P., 2016. The WITCH 2016 Model - Documentation and Implementation of the Shared Socioeconomic Pathways, FEEM Working Paper No. 2016.42, Fondazione Eni Enrico Mattei.

Eyckmans, J., and H. Tulkens, 2003, Simulating Coalitionally Stable Burden Sharing Agreements for the Climate Change Problem. Resource and Energy Economics 25, no. 4 (October 2003): 299–327.

Fankhauser, S., 1995. Valuing climate change: the economics of the greenhouse. Routledge

Kornek, U., Steckel, J.C., Lessmann, K., Edenhofer, E. 2017. The climate rent curse: new challenges for burden sharing. International Environmental Agreements 17 (6), 855–882.

Kriegler, E., Hall, J.W., Held, H., Dawson, R., Schellnhuber, H.J., 2009. Imprecise probability assessment of tipping points in the climate system. PNAS 106, 5041–5046.

Lempert, R.J., Sanstad A.H., and M.E. Schlesinger. 2006. Multiple Equilibria in a Stochastic Implementation of DICE with Abrupt Climate Change. Energy Economics 28, no. 5 (November 1, 2006): 677–89.

Luderer, G., Pietzcker R.C., Bertram, C., Kriegler, E., Meinshausen, M., Edenhofer, O., 2013. Economic mitigation challenges: how further delay closes the door for achieving climate targets. Environmental Research Letters 8(3), 034033.

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

Rinne, H., 2014. The Hazard rate: Theory and inference (with supplementary MATLAB-Programs).