Technical Appendix

This appendix provides a complete technical description of the model expressed through equations and pseudocode, plus supplementary information that would not fit in the main manuscript. The model code is available from the authors upon request.

1. Model Parameters and Variables

Table A.1. Model parameters and variables

Variable or parameter	Explanation	Measurement units
	Global input parameters	
CGB scale	Scaling factor for exclusive club good. See Table 1 (in the main text) for numerical values.	-
Vulnerability Weight	Degree of differentiation between more and less vulnerable countries. See Table 1 for numerical values.	-
Global Damage	The difference in global climate damage costs between the no-club scenario and the scenario where *all* actors spend 1% of their GDP to mitigate climate change. See Table 1 for numerical values.	% of Gross Global Product (GGP)
Club fee	The mitigation expenditure required by club members. Set to 1% of entrant's GDP.	% of GDP
Conditional commitments	Conditional commitments allowed?	Yes/No
	Agent-specific input parameters	
GDP_i	GDP at market exchange rates, 2013 (World Bank 2014).	Share of GGP
$Emissions_i$	Emissions from fossil fuels and cement, 2013 (Global Carbon Project 2014) + Net emissions from land-use change and forestry, 2011 (World Resources Institute 2014).	Share of global emissions
$NDGAIN_i$	Vulnerability Scores, 2012 (Notre Dame Global Adaptation Index 2014). $\in[0,1]$	-
$Enthusiast_i$	Is *i* "enthusiastic"?	Yes/No
	Agent-specific variables derived in the model	
$Vulnerability_i$	The percentage loss in GDP_i arising when the global loss is 1% of GGP. $$\frac{NDGAIN_i + (Vulnerability\ Weight - 1) \times (NDGAIN_i - \overline{NDGAIN})}{\overline{NDGAIN}}$$ $\overline{NDGAIN}$ is GDP-weighted mean *NDGAIN.*	$\frac{\% \text{ of } GDP_i}{\% \text{ of } GGP}$
$Expenditure_i$	Member *i*'s expenditure on mitigation. Without conditional commitments, it is always equal to *Club fee*. A conditional commitment raises it above *Club fee*.	% of GDP_i
Benefit from $expenditure_{ji}$	The benefit to *j* of *i*'s mitigation expenditure $$Global\ Damage \times Vulnerability_j \times Emissions_i \times \sqrt{Expenditure_i}$$ Implies that domestic expenditure (abatement cost) is a quadratic function of mitigation. Benefits are proportional to *i*'s share of global emissions.	% of GDP_j
Marginal benefit from $expenditure_{ji}$	The first derivative of the above with respect to $Expenditure_i$ $$\frac{Global\ Damage \times Vulnerability_j \times Emissions_i}{2 \times \sqrt{Expenditure_i}}$$	% of GDP_j
Club $benefit_i$	The exclusive club-good benefit earned by member *i.*	% of GDP_i

	$CGB\ scale \times \ln\left(\sum_{m \neq i}^{members} GDP_m\right)$	
Benefit of entry$_i$	The private benefit of becoming a member equals club benefits plus the privately captured benefits from own mitigation expenditure of 1% of GDP (in terms of damage costs avoided). $Club\ benefit_i + (Benefit\ from\ expenditure_{ii} \vert Expenditure_i = 1)$	% of GDP_i
Club benefit(+j)$_i$	The exclusive club good benefit earned by member *i* if non-member *j* joins. $CGB\ scale \times \ln\left(\sum_{m \neq i}^{members} GDP_m + GDP_j\right)$	% of GDP_i
Benefit from expansion$_{ji}$	The benefit to member *i* if non-member *j* joins equals the increase in the club-good benefit plus *i*'s benefit from *j*'s mitigation. $Club\ benefit(+j)_i - Club\ benefit_i + (Benefit\ from\ expenditure_{ij} \vert Expenditure_j = 1)$	% of GDP_i
WTP$_{ij}$	Additional expenditure *i* is willing to undertake to induce *j* to join $\frac{Benefit\ from\ expansion_{ji}}{1 - Marginal\ benefit\ from\ expenditure_{ii}}$ The denominator captures that *i* will enjoy a marginal benefit from its own additional effort, which increases its willingness to make such efforts.	% of GDP_i
Benefit from conditional commitments$_j$	The sum of benefits to *j* from members' *WTP*$_{ij}$ $\sum_{i=1}^{Members} WTP_{ij} \times Marginal\ benefit\ from\ expenditure_{ji}$	% of GDP_j
Ratio$_j$	The ratio between *Benefit from conditional commitments*$_j$ and *j*'s net cost of joining $\frac{Benefit\ from\ conditional\ commitments_j}{Club\ fee - Benefit\ of\ entry_j}$	-
Payoff$_i$	*Payoff* incorporates (i) *Club benefit*$_i$, (ii) *Expenditure*$_i$ on mitigation, and (iii) the benefit to *i* from the club's mitigation. For non-members, (i) and (ii) both equal zero. Payoff is normalized to be zero in the no-club scenario. $Club\ benefit_i - Expenditure_i + \sum_{j=1}^{members} Benefit\ from\ expenditure_{ij}$	% of GDP_i

2. Model pseudocode

Initialization

```
Each agent i:
   Calculate Vulnerability_i
   Calculate Damage cost_i
```

Execution

Step 1

```
Each enthusiast:
   Become member
Loop:
   Each non-member i:
      Calculate Benefit of entry_i
      Agent with greatest Benefit of entry:
         If Benefit of entry_i > Club fee:
            Become member
                  Else:
            End loop
```

Step 2 (applicable only if conditional commitments are allowed)

```
Loop:
   Each non-member j:
      Calculate Benefit of entry_i
      If Benefit of entry_i > Club fee[1]:
         Become member
      Else:
         Ask each member i:
            Calculate WTP_ij
         Calculate Ratio_j
   Non-member with the largest Ratio_j:
      If Ratio_j > 1:
         Become member
         Ask other members i:
            Increase Expenditure_i by WTP_ij/Ratio_j[2]
      Else: End loop
```

Step 3

```
Loop until no more enthusiasts want to leave the club:
   Each enthusiast i:
      Calculate Payoff_i:
      If Payoff_i < 0:
         Become non-member
   If any enthusiasts left the club:
      Each reluctant member i:
```

[1] This inequality could in theory hold when another country has entered because of conditional commitments and hence enlarged the club-good benefit since the last time the country concerned declined to join.

[2] The club's additional mitigation is distributed in proportion to WTP_{im}. Dividing by $Ratio_j$ implies that the members undertake the additional effort necessary to make *j*'s net cost of entry zero.

```
      Calculate Benefit of entry_i
      If Benefit of entry_i < Club fee:
         Become non-member
    Repeat Step 2
  Else: model stops
```
[3]

3. On sequencing, stochasticity, and stability

In agent-based models (ABMs), a command is executed by agents sequentially (i.e., one agent at a time). The default sequence is random order. In this model, the order in which non-members negotiate with the group can affect outcomes. It seems more realistic that the club will prioritize negotiating with the most likely candidate than that it will choose a random candidate. We therefore code negotiations as the following two-step loop: 1) Identify the candidate with the most favorable benefit-cost ratio for membership 2) Negotiate with that candidate. A negotiation round is defined as one such loop. Step 1 is repeated if and only if Step 2 leads to club expansion. The number of negotiation rounds before membership stabilizes is therefore generally equal to the (equilibrium) number of reluctant actors recruited.

The above procedure effectively makes the model deterministic, eliminating the need for multiple runs for each input vector.

4. List of vulnerable countries

Table A2. The 30 most vulnerable countries included in the model, according to the Notre Dame Global Adaptation Index 2014 (vulnerability scores for 2012). Listed in descending order of vulnerability, by row.

Burundi	Mali	Yemen, Rep.
Sierra Leone	Solomon Islands	Uganda
Afghanistan	Madagascar	Rwanda
Central African Republic	Tanzania	Benin
Togo	Haiti	Angola
Liberia	Guinea-Bissau	Mozambique
Congo, Dem. Rep.	Burkina Faso	Nigeria
Ethiopia	Kenya	Cote d'Ivoire
Guinea	Niger	Papua New Guinea
Chad	Sudan	Cambodia

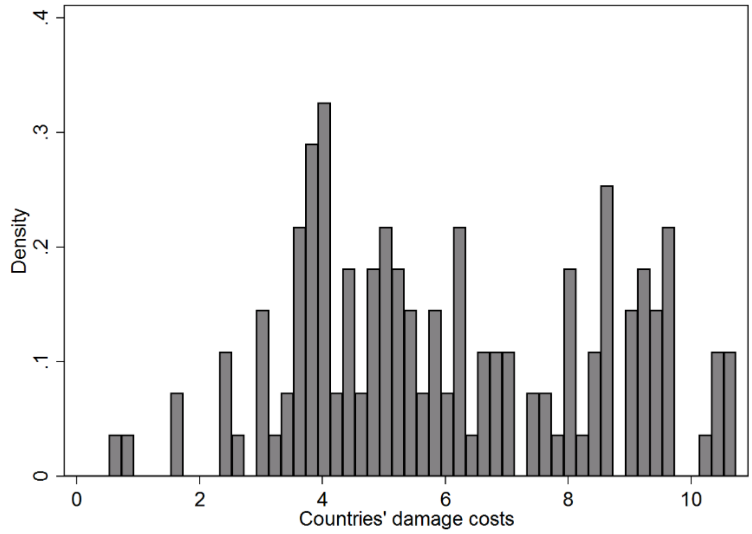
5. Sensitivity analysis

Damage cost magnitude and distribution

Figures A1 and A2 report the sensitivity of our results concerning the magnitude and distribution of damage costs. *Global Damage Costs* are varied by +/- 50% and *Vulnerability weight* is changed from the baseline value of unity to zero (i.e., damages are proportional to GDP as *Vulnerability*$_i$ equals 1) and 2 (i.e., damage costs are distributed with larger variance, as shown in Figure A1). To keep the number of scenarios manageable, we limit the analysis to four *CGB scale* values (0, 0.1, 0,2, and 0.25).

[3] New negotiations start between any remaining members and non-members.

Figure A1. Histogram of actors' BAU damage costs (% of GDP) when *Vulnerability weight* equals 2 and *Global damage cost* equals 3% of GGP.



Reducing the club's effect on climate damage, so that spending 1% of GGP yields benefits of only 1.5% (rather than 3%) of GGP, entails that no coalition persists for a zero or only a small (0.1) club good, with or without conditional commitments.

Figure A2. Sensitivity of simulated participation (% of global emissions) to *Global Damage Costs* and *Vulnerability weight.* With conditional commitments.

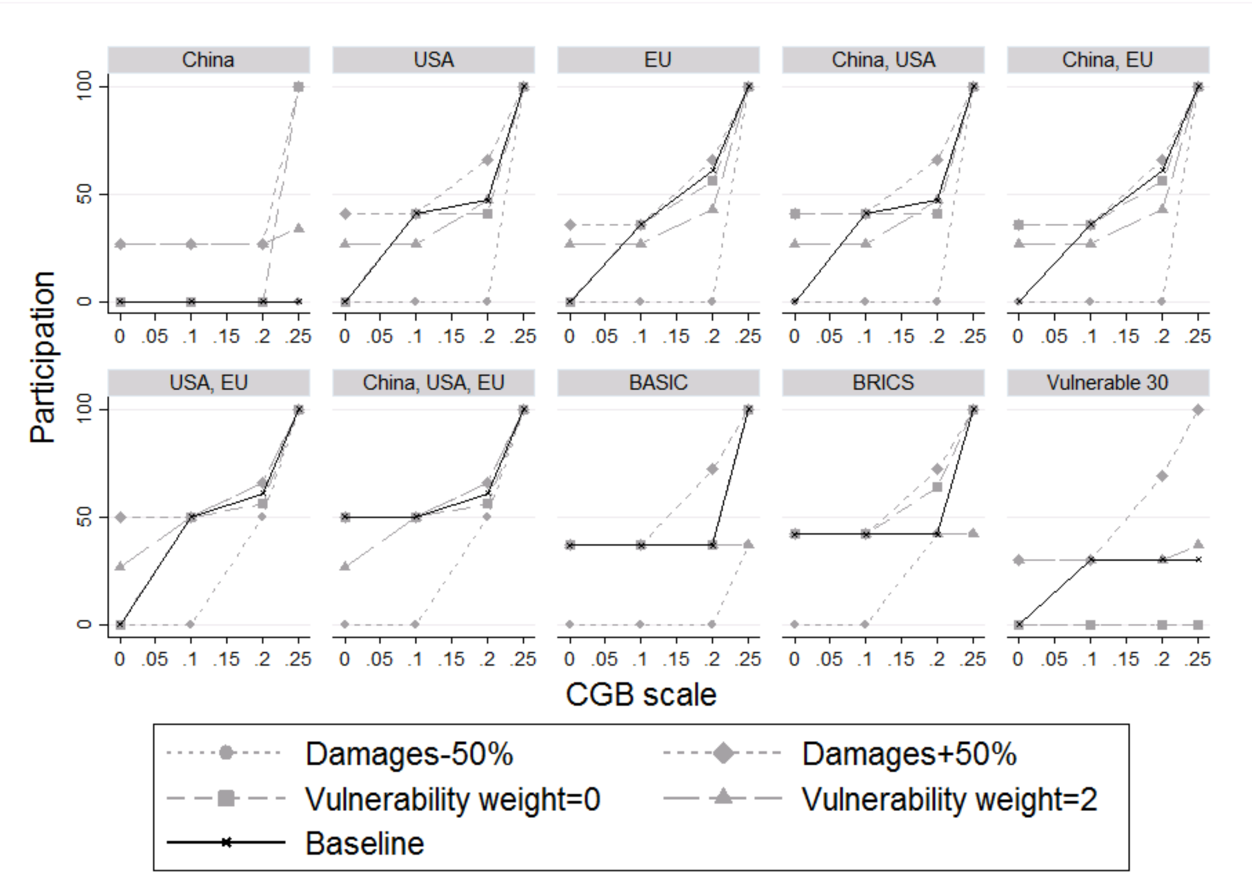
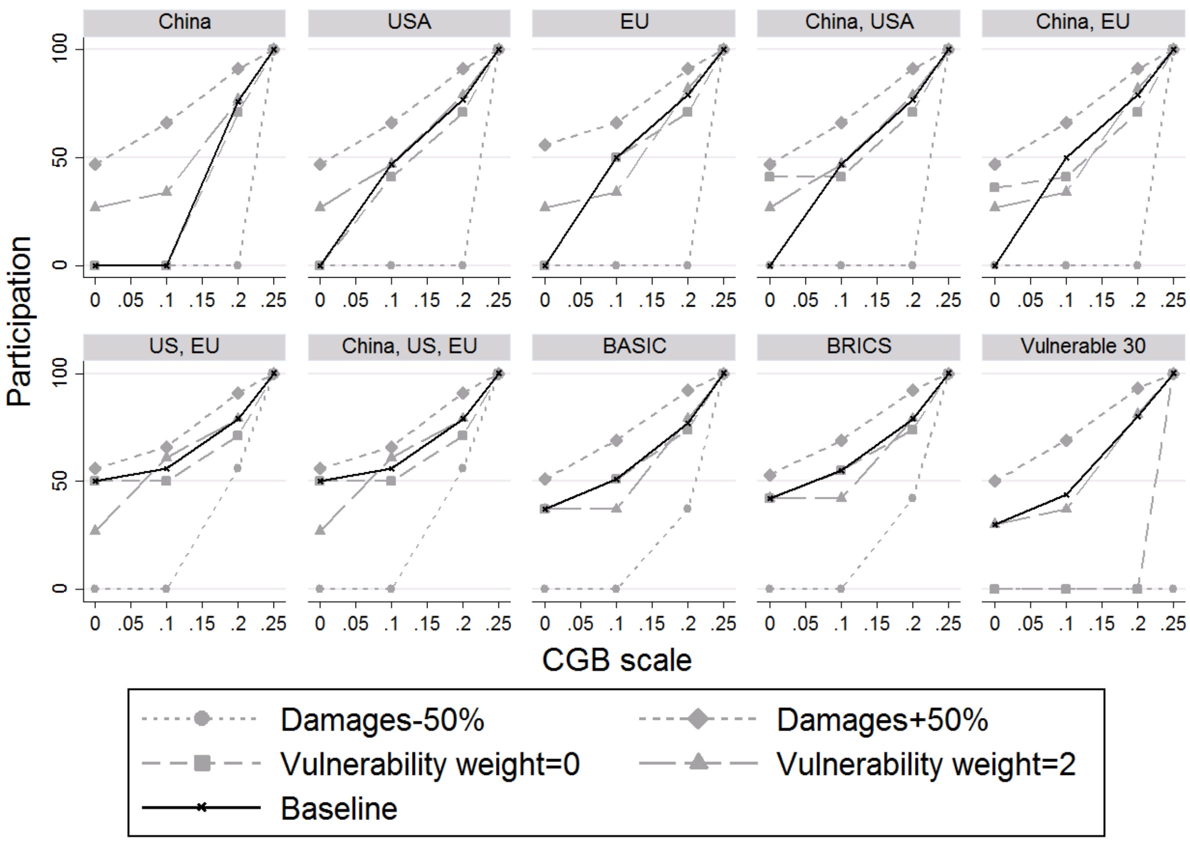


Figure A3. Sensitivity of simulated participation (% of global emissions) to *Global Damage Costs* and *Vulnerability weight.* With conditional commitments.



With an intermediate club good (0.2), only three coalitions persist absent conditional commitments. Conditional commitments facilitate one additional coalition, while enlarging two that would persist even without conditional commitments.

For the largest club-good size (0.25), the reduction in *Global Damage Costs* has less effect, although it reduces participation in two clubs without conditional commitments and eliminates the club based on the Vulnerable 30 with and without conditional commitments. The Vulnerable 30 coalition is hence unable to initiate a club in any scenario with the reduced environmental benefits. This result is due to China's reduced incentive for mitigation. In sum, when the *Global damage cost* is small (1.5%), a large club-good benefit is often a necessary condition for cooperation.

Conversely, *increasing* the club's potential climatic benefits from 3% to 4.5% of GGP entails three distinct effects. First, participation increases; indeed, one actor (China) now even has a purely selfish incentive for acting *unilaterally*.[4] Second, the conditional commitments' leverage increases; indeed, they now make a difference in most cases – including all cases with zero club-good benefits. An

[4] This is a rare instance where clubs can emerge absent enthusiasts. With conditional commitments, participation rates reach 47%, 66%, 91%, and 100% under the respective club-good benefit sizes. Without conditional commitments, other countries join only for the largest club-good benefit size, in which case all join.

exception concerns the most optimistic club-good benefit scale assumed (0.25), which mostly generates universal participation even *without* conditional commitments. Finally, the impact of the club-good benefits declines somewhat. The reason is that the increased environmental benefit ensures considerable participation even with small (or even zero) club-good benefits.

Varying the value of *Vulnerability weight* results in smaller changes than those obtained when varying *Global damage cost*, indicating that for participation rates, the mean damage cost matters more than the variance. Changing *Vulnerability weight* has a non-linear effect on participation and the sign of this effect depends on who the enthusiasts are. If damage costs are assumed to be proportional to GDP (*Vulnerability weight* = 0), the mitigation incentives of the European Union and the United States increase (relative to our baseline scenario), while the mitigation incentives of emerging economies decrease. Stronger mitigation incentives increase reluctant actors' likelihood of joining the club. Assuming a more unequal distribution of vulnerability than in the baseline scenario (*Vulnerability weight* = 2) changes relative incentives in the opposite direction, giving China (once again) an incentive for unilateral action because by spending 1% of its GDP on mitigation, it avoids damages to itself worth more than 1% of its GDP.[5]

Overall, the vulnerability distribution has systematic effects on who participates but not on total participation. The vulnerability distribution shows no systematic interaction with the effectiveness of club-good benefits or of conditional commitments.

Trade relations

To incorporate trade relations among the top 10 emitters, we assume that the benefit to two such actors in a club together is a function of the amount these two countries trade. Because trade data is included for only the top 10 emitters, the benefit of being in a club with any other actor is a function of that actor's GDP, as before.

Let *Mt10* be the set of members that are among the top 10 emitters of GHGs (see Table 2), let *Mr* be the other members, and let *n* be the set of all actors. Subscript *ij* denotes a trade flow between countries *i* and *j.* For $i \in Mt10$, Equation 2 is changed to

$$Club\ benefit_i = \text{CGB scale} \times \ln\left(\frac{\sum_{j=1}^{Mt10}(exports_{ij} + imports_{ij})}{\sum_{j=1}^{n}(exports_{ij} + imports_{ij})} + \sum_{j \neq i}^{Mr} GDP_j\right)$$

As before, *GDP* is measured as a share of GGP. For $i \in Mr$, Equation 2 in the main text remains unchanged.

Data on trade flows involving the European Union have been obtained from the European Commission.[6] All other data are from the OECD.[7] All data are from 2013, with the exception that the most recent data available on Iran is from 2011.

[5] This is another instance where clubs can emerge absent enthusiasts. With conditional commitments, participation rates reach 27% (China alone), 34%, 77%, and 100% under the respective club-good benefit sizes. Without conditional commitments, only India joins (and it does so only for the largest club-good benefit).

[6] European Commission 2015.

[7] OECD 2015.

Affinity and hostility (conflict)

The assumption underlying the tests that include affinity and hostility (conflict) is that countries tend to treat friends better than enemies, everything else constant. In a climate club context, this inclination may be expressed in an actor's tolerance of another actor's free riding and/or in its attitude towards differentiation of membership terms. For both dimensions we assume that a country will have a certain zone of indifference, implying that the behavior hypothesized will be found only when affinity/hostility scores *exceed* a certain threshold (see specifications below).

Affinity

To measure affinity, we use Voeten et al.'s UN General Assembly voting similarity index,[8] which is defined as

$$\frac{total\ number\ of\ votes\ where\ both\ states\ agree}{total\ number\ of\ joint\ votes}$$

We use data from the 10 most recent years available (2005–2014), taking the average index scores over this period. Because we treat the European Union as a single actor, and we calculate the (unweighted) average scores across its members. For every actor, we use the average voting similarity with the world (all countries except itself) as a benchmark.

Table A3 lists, for each top 10 emitter, the other top 10 emitters that vote most similarly. For the modelling, we include up to three relations for each actor. We require, as a further criterion for inclusion, that the voting similarity must be at least 10% above the actor's average voting similarity with the world. Relations that fail this criterion are displayed in parentheses. Twenty-one affinity relations are hence included in the analysis. We weight all these relations equally, and assume that there is a political benefit of club membership equal to .1% of GDP per "friend" that is also a member. Hence, *i*'s affinity benefit of being in a club with *j* – denoted $Affinity_{ij}$ – takes the values zero and .1.

The .1% figure is set somewhat arbitrarily and is difficult to ground truth. We deem the non-material benefits created by affinity to be an order of magnitude smaller than the material costs of reducing emissions that club members must shoulder (1% of GDP). We set the figure equal to the upper limit of what we see as plausible values, because the purpose is to test our main results' sensitivity to including affinity relations.

The identified relationships largely mirror the general political divide between the G77 and the OECD countries. Historically, this divide has been very important in climate negotiations.

Table A3. High affinity relationships within our sample of main actors, based on United Nations General Assembly (UNGA) voting records

Main actor ↓	Highest affinity score	Second highest	Third highest
Brazil	(Indonesia)	(China)	(Iran)
Canada	EU	Japan	US
China	Iran	Indonesia	Brazil
EU	Japan	(Canada)	(Brazil)
India	China	Iran	Indonesia

[8] Voeten et al. 2009. We use the version that includes abstention from voting, counting it as half-agreement with a yes or no vote.

Indonesia	Iran	China	Brazil
Iran	China	Indonesia	Brazil
Japan	EU	(Brazil)	(Canada)
Russia	China	(Iran)	(Indonesia)
US	Canada	EU	Japan

Note: Parentheses indicate that the score fails to meet the > 10% requirement.

Hostility (conflict)

UNGA voting records are sometimes used to measure hostility or conflict levels as well. Doing so here would, however, have blurred the distinction between divergent voting expressing different overall political alignments and divergence expressing truly severe (bilateral) conflict.[9] Since the latter is the more important potential source of "disturbance" for our analysis, we have based our estimates of hostility levels on data identifying and describing *severe* conflict, at the level of "militarized interstate disputes" (MIDs).[10] To qualify as a conflict for this robustness test, we require that two top 10 emitters have been opponents in at least two MIDs during the period 2001–2010. An overview of such conflicts is displayed in Table 2.

Table A4. Conflict relationships within our sample of main actors, based on MIDA_4.01 and MIDB_4.01

"Offensive" actor	Initiator (number of disputes)	Hostility level	Opponent	Initiator (number of disputes)	Hostility level
China	5	2.7	Japan	2	2.1
China	5	2.8	USA	0	2.2
Russia	5	3.4	Japan	0	2.2
USA	4	3.4	Iran	3	2.3
India	2	2.3	China	1	1.7

Table A4 shows that some of these conflicts are *asymmetric* in at least two respects: one of the parties stands out as the initiator of the dispute, and one of the parties (usually the initiator) has higher scores on "hostility level" (~ militarization) of the dispute.[11] Other things being equal, we may expect a party's offensiveness in a particular MID to reflect (a) its interest in having the opponent change a certain policy or behavior, and (b) a perception of being sufficiently powerful to persuade or coerce the opponent to do so. To the extent that these interests and perceptions "spill over" to international climate change politics, we would expect the asymmetry in MID behavior to be *reflected* in climate club considerations. It may, however, also be *softened* by functional and/or ideological distance between the MID dispute(s) and the climate policy domain.[12] To translate this

[9] To illustrate, the voting similarity score between the United States and Brazil is almost as low (.25) as that between the United States and Iran (.18).

[10] Palmer et al. 2015. Data available from http://cow.dss.ucdavis.edu/data-sets/MIDs. Since the Dyadic MID Data File has not been updated beyond 2001 we have ourselves extracted information about bilateral relations from the MID files referring to "disputes" and "participants."

[11] Below we refer to the combination of these two dimensions as "offensiveness."

[12] For illustrative empirical evidence from another setting, see for example, Kohl and Randall 1991.

line of reasoning into a template for differentiating offensiveness scores we have used a two-step procedure:

Step 1: Combining initiator role and level of hostility into an index of *"offensiveness"*

Aspect 1: *Initiator* (counting and weighting a party's number of MID initiating roles):

0 = 0; 1–2 = 1; 3–4 = 2; ≥ 5 = 3.

+

Aspect 2: *Level of hostility* (average for all parties 2001–2010 = 2.8):

≤2.3 = 1; 2.4–3.2 = 2; ≥ 3.3 = 3.

Step 2: Specify *likely impact of offensiveness scores on preferences* (expressed in terms of change in our current GDP measures).

Because we have a moderate number of conflicts and actors to consider, we merge offensiveness scores into a dichotomous distinction between a "low" level of conflict (offensive scores 1–3) and a "high" level of conflict (offensive scores 4–6). In the model runs, "low" = 1 and "high" = 2. Table A5 displays, for each conflict in our sample, the offensiveness score and the conflict level. We assign level 2 conflicts 50% higher weight (importance) than level 1 conflicts. Because we have no basis for assuming that conflict is inherently more important than affinity (or vice versa), and because the two are measured in different terms, we balance the two by requiring that the weighted sums of conflicts and affinities cancel out. This balance is achieved by setting the political cost of co-membership with the opponent of a level 1 (2) conflict to .16% (.24%) GDP. Hence, *i*'s conflict cost of being in a club with *j* – denoted $Conflict_{ij}$ – takes the values zero, .16, or .24.

While several other approaches might be equally valid, the one outlined here facilitates producing a general impression of the effect of conflict on climate clubs.

Table A5. Offensiveness scores and conflict levels of MIDs among the top 10 emitters of greenhouse gases

	Offensive party			Opponent (defensive party)	
Country	Offensiveness	Conflict level	Country	Offensiveness	Conflict level
China	5	2	Japan	2	1
China	5	2	USA	0	1
Russia	6	2	Japan	1	1
USA	5	2	Iran	3	1
India	3	1	China	2	1

Inclusion of affinity and conflict in the model code

The following modifications are made to equations from Table A1, with new terms in bold:

$$Benefit\ of\ entry_i = Club\ benefit_i + (Benefit\ from\ expenditure_{ii} | Expenditure_i = 1) + \boldsymbol{\sum_{j \neq i}^{members} (Affinity_{ij} - Conflict_{ij})}$$

$$Benefit\ from\ expansion_{ij} = Club\ Benefit(+j)_i - Club\ Benefit_i + (Benefit\ from\ expenditure_{ij} | Expenditure_j = 1) + \boldsymbol{Affinity_{ij}\text{-}Conflict_{ij}}$$

$$Payoff_j = Club\ benefit_i - Expenditure_i + \sum_{j=1}^{members} Benefit\ from\ expenditure_{ij} + \sum_{j \neq i}^{members} \boldsymbol{(Affinity_{ij}\text{-}Conflict_{ij})}$$

6. References

European Commission 2015. Market Access Database. Available from http://madb.europa.eu/madb/statistical_form.htm, accessed 6 October 2015.

Kohl, Wilfred L. and Carol W. Randall. 1991. OPEC and the World Oil Market: The March 1983 London Agreement. Washington D.C.: Georgetown University, School of Foreign Service, Institute for the Study of Diplomacy. Case 123.

OECD 2015. STAN Bilateral Trade in Goods by Industry and End-use, ISIC Rev.4. Available from http://stats.oecd.org/Index.aspx?DataSetCode=BTDIXE_I4, accessed 6 October 2015.

Palmer, Glenn, Vito d'Orazio, Michael Kenwick, and Matthew Lane. 2015. The MID4 dataset, 2002–2010: Procedures, coding rules and description. *Conflict Management and Peace Science* 32 (2):222–242.

Voeten, Erik, Anton Strezhnev, and Michael Bailey. 2009. "United Nations General Assembly Voting Data", Harvard Dataverse, V11. Available from http://hdl.handle.net/1902.1/12379