Uncertainty, Learning and International Environmental Agreements – The Role of Risk Aversion

Alistair Ulph
(University of Manchester)

Pedro Pintassilgo
(University of Algarve)

and

Michael Finus
(University of Bath)

Abstract

This paper analyses the formation of international environmental agreements (IEAs) under uncertainty, learning and risk aversion. It bridges two strands of the IEA literature: (i) the role of learning when countries are risk neutral; (ii) the role of risk aversion under no learning. Combining learning and risk aversion seems appropriate as the uncertainties surrounding many international environmental problems are large, often highly correlated (e.g. climate change), but are gradually reduced over time through learning. The paper analyses three scenarios of learning. A key finding is that risk aversion can change the ranking of these three scenarios of learning in terms of welfare and membership. In particular, the negative conclusion about the role of learning in a strategic context under risk neutrality is qualified. When countries are significantly risk averse, then it pays them to wait until uncertainties have been largely resolved before joining an IEA. This may suggest why it has been so difficult to reach an effective climate change agreement.

JEL-Classification: C72, D62, D80, Q54

Keywords: international environmental agreements, uncertainty, learning and risk aversion, game theory

We are grateful to Larry Karp for helpful comments on an earlier version of this paper. The usual disclaimer applies.

1. Introduction

Environmental issues such as climate change pose three key challenges for economic analysis: (i) there are considerable uncertainties about the likely future costs of environmental damages and abatement; (ii) our understanding of these uncertainties changes over time as a result of learning more about climate science, possible technological responses and behavioral responses by households, firms and governments; (iii) the problem is global, but since there is no single global agency to tackle climate change, policies need to be negotiated through international environmental agreements (IEAs).[1,2] Recently, these three issues have begun to be integrated in one framework. Two stands of literature can be distinguished.

The first strand of literature studies uncertainty and IEA formation with the focus on the role of *learning*, but under the assumption of risk neutrality. Ulph and Ulph (1996) and Ulph and Maddison (1997) compare the fully cooperative and the non-cooperative scenarios when countries face uncertainty about damage costs. They show that the value of learning about damage costs may be negative when countries act non-cooperatively and damage costs are correlated across countries. Na and Shin (1998), Ulph (2004), Kolstad (2007), Kolstad and Ulph (2008, 2011) have considered how the prospect of future resolution of uncertainty affects the incentives for countries to join an IEA. Kolstad and Ulph consider a model where countries face common uncertainty about the level of environmental damage costs.[3] Three scenarios of learning are considered: with *full learning*, uncertainty about damage costs is resolved before countries decide whether to join an IEA; with *partial learning*, uncertainty is resolved after countries decide whether to join an IEA, but before they choose their emissions levels; with *no learning*, uncertainty is neither resolved before stage 1 nor stage 2. They showed that the prospect of learning, either full or partial, generally reduces the expected total

---

1 On the first two issues, see for instance Arrow and Fisher (1974), Epstein (1980), Kolstad (1996a,b), Ulph and Ulph (1997), Gollier, Julien and Treich (2000) as well as Narain, Fisher and Hanemann (2007).

2 On the third issue, see for instance the classic papers by Carraro and Siniscalco (1993) and Barrett (1994). The most influential papers have been collected in a volume by Finus and Caparros (2015) who also provide a survey.

3 By common uncertainty we mean that each country faces the same *ex-ante* distribution of possible damage costs, and when uncertainty is fully resolved they face the same *ex-post* level of damage costs, i.e. the risks they face are fully correlated across countries. Kolstad and Ulph (2011) extend this model to consider the case where the risks each country faces are uncorrelated. Uncorrelated uncertainty is also considered in a slightly different model in Finus and Pintassilgo (2013) and empirically investigated in a climate model with twelve world regions in Dellink and Finus (2012).

payoff in stable IEAs. In particular, Kolstad and Ulph (2008, 2011) showed that *partial learning* would yield the highest total payoff for only a small proportion of parameter values. For a significant majority of parameter values, the highest expected total payoff arose under *no learning*. Hence, it is better to form an IEA before waiting for better information: removing the "veil of uncertainty" seems to be detrimental to the success of international environmental cooperation.

All these models have assumed that countries are risk neutral. However, in the climate context, risks are highly correlated and hence possibilities for risk sharing are limited so that the assumption of risk aversion may be quite relevant. Therefore, we extend the two-stage coalition formation setting by Kolstad and Ulph (2008) by departing from the assumption of risk neutrality. In this paper, we allow for countries to be risk averse, and show that if countries have a relatively high degree of risk aversion, then for a majority of parameter values *full learning* yields higher expected total utility than *no learning*. This may help to explain why it has taken such a long time between the start of the process of tackling climate change (the Kyoto Protocol) to reach a more substantial agreement in Paris – countries are risk averse and so needed to wait until they had more information about the risk of climate change before committing to significant action to tackle climate change.

The second strand of literature studies uncertainty and IEA formation with the focus on the role of *risk aversion*, though under the assumption of no learning. Endres and Ohl (2003) show in a simple two-player prisoners' dilemma, using the mean-standard deviation approach to capture risk aversion, that risk aversion can increase the prospects of cooperation once it reaches a certain threshold. The reason is that the benefits of mutual cooperation increase relative to the payoffs of unilateral cooperation and no cooperation because cooperation reduces the variance of payoffs. The more risk averse players are, the more attractive cooperation becomes compared to free-riding. In their model, there is a first threshold above which the prisoners' dilemma turns into a chicken game and a second threshold above which the game turns into an assurance game. Compared to their paper, we allow for an arbitrary number of players, model cooperation as a two-stage coalition formation game and consider explicitly the role of learning.

Bramoullé and Treich (2009) consider risk-averse players in a global emission model, in which all players behave non-cooperatively as singletons. They show that equilibrium emissions are lower under uncertainty than under certainty, as part of a hedging strategy, but

the effect on global welfare is ambiguous. The authors also find that emissions decrease with the level of risk aversion. Unlike our paper, Bramoullé and Treich are not concerned with learning and coalition formation.

Boucher and Bramoullé (2010) consider the effects of risk aversion on coalition formation, but only with no learning. They analyze the formation of an international environmental treaty using a similar coalition game and payoff function as adopted in this paper. Using an expected utility approach, their analysis focuses on the effect of uncertainty and risk aversion on signatories' efforts, the participation level in an agreement and total expected utility. They show that if additional abatement reduces the variance of countries' payoffs, then, under risk aversion, an increase in uncertainty tends to increase abatement levels and may decrease equilibrium IEA membership while the reverse is true if additional abatement increases the variance of countries' payoffs.[4] In this paper, our model of no learning satisfies the first condition, but we extend the analysis of Boucher and Bramoulle (2010) by considering also partial learning and full learning.

Thus, taken together, in our paper, we generalize the analysis of Kolstad and Ulph (2008) by allowing for risk aversion, and the analysis of Boucher and Bramoulle (2010) and Endres and Ohl (2003) by considering the role of learning. The key findings are that as countries become more risk averse it is no longer the case that for most parameter values the scenario of "No Learning" yields the highest expected aggregate utility, but increasingly it is the scenario "Full Learning". Moreover, the set of parameter values for which the scenario "Partial Learning" yields the highest expected aggregate utility, which is a small subset of such values when countries are risk neutral, becomes even smaller as countries become more risk averse. Thus, we qualify the negative conclusion about the role of learning in a strategic context if players are sufficiently risk averse.

In our model, emissions last for just one period, which may seem restrictive in the context of dynamic environmental problems such as climate change. However, it has been shown in the literature on IEA formation under uncertainty that one period models produce similar results to multi-period models. For instance, Kolstad and Ulph (2008), using a one-period model, and

4 Hong and Karp (2013) show that it does not matter whether one analyses the provision of a public good or the amelioration of a public bad. What matters is whether players' actions increase or decrease the volatility of payoffs. In our model, as in Endres and Ohl (2003) and the emission game in Boucher and Bramoulle (2010), abatement (emissions) reduces (increase) the volatility of payoffs.

Ulph (2004), using a two-period model, found similar results regarding the implications of different scenarios of learning for the size of an IEA and expected welfare.[5]

Taken together, the tension we seek to capture in our modeling is between No Learning where countries decide to join an IEA and base their decisions for ever on expected damage costs ignoring any later information, Full Information where countries delay making any decision to join an agreement until (almost) all uncertainty about damage costs has been resolved, or Partial Learning where countries start the process of joining an IEA knowing that as they get better information they will be able to use that to adjust their emissions policies. This would seem to be particularly relevant to the kind of situation to which Weitzman (2009) has drawn attention – a small probability of catastrophic climate change.

The paper proceeds as follows. In section 2, we set out the theoretical model and present our theoretical results in section 3. Section 4 presents some simulation results while Section 5 summarizes our main conclusions and implications for future research.

2. The Model

2.1 No Uncertainty

To establish the basic framework, we set out the model with no uncertainty. There are N identical countries, indexed $i = 1, ..., N$. Country i produces emissions x_i with aggregate emissions denoted by $X = \sum_{i=1}^{N} x_i$. Aggregate emissions cause global environmental damages. The cost of environmental damages per unit of global emissions is γ and the benefit per unit of individual emissions is normalized to 1. (Thus, γ essentially measures the cost-benefit ratio.) The payoff to country i, as a function of own and aggregate emissions, is given by

$$\pi_i(x_i, X) \equiv \pi_0 + x_i - \gamma X \tag{1}$$

[5] In many papers with a dynamic payoff structure but fixed membership, results are qualitatively similar to the one period emission game (e.g. Rubio and Casino 2005). The extension to flexible membership would be more interesting but is technically very challenging. See Rubio and Ulph (2007).

with π_0 a positive constant. In this simple model with a linear payoff function, following the literature, the (continuous) strategy space can be normalized to $x_i \in [0,1]$.[6] To make this model interesting, we make the following assumption:

Assumption 1: i) $1/N < \gamma < 1$; (ii) $\pi_0 \geq N-1$.

The individual benefit exceeds the individual unit damage cost from pollution, i.e. $1 > \gamma$ (so countries pollute in the Nash equilibrium) but does not exceed the global unit damage cost, i.e. $1 < \gamma N$ (so countries abate in the social optimum); the second condition is a sufficient condition such that $\pi_i(.) \geq 0$, which we will need when we consider expected utility.[7]

In order to study coalition formation, we employ the widely used two-stage model of IEA formation (Carraro and Siniscalco, 1993 and Barrett, 1994) which is solved backwards. In stage 2, the *emission game*, for any arbitrary number of IEA members n, $1 \leq n \leq N$, the members of the IEA (which we denote by the symbol *c* for *c*oalition countries) and the remaining countries (which we denote by the symbol *f* for *f*ringe countries) set their emission levels as the outcome of a Nash game between the coalition and the fringe countries.[8] That is, the coalition members together maximize the aggregate payoff to their coalition, whereas fringe countries maximize their own individual payoff. Given $1 > \gamma$, $x_f = 1$ follows; coalition members chose $x_c = 0$ if $1 \leq \gamma n$, and so $\pi_f(n) = \pi_0 + 1 - \gamma(N-n)$ and $\pi_c(n) = \pi_0 - \gamma(N-n)$; however if $1 > \gamma n$, then coalition members will also pollute, $x_c = 1$ and so $\pi_c(n) = \pi_f(n) = \pi_0 + 1 - \gamma N$.[9]

Knowing the payoffs to coalition and fringe countries for any arbitrary number of IEA members, we then determine the stable (Nash) equilibrium in stage 1, the *membership game*.

---

6 Either benefits from emission are lower than damage costs in which case equilibrium emissions are zero, or the reverse is true in which case equilibrium emissions would obtain their maximum. Thus, the upper bound is set to 1 here.

7 The lowest possible payoff is derived if a country abates and all other countries free-ride, which is given by $\pi_i = \pi_0 + 1 - \gamma(N-1)$ and hence, given that $\gamma < 1$, $\pi_0 \geq N-1$ is a sufficient condition for $\pi_i(.) \geq 0$.

8 A sequential Stackelberg game in the second stage, as an alternative assumption (e.g. Barrett 1994), would make no difference here as players have dominant strategies. This also applies to Boucher and Bramoullé (2010).

9 It is now evident why we need Assumption 1: it avoids trivial outcomes where all countries either abate or pollute in the Nash equilibrium and the social optimum (and hence also for partial cooperation).

No member should have an incentive to leave the coalition, e.g. the coalition is internally stable, $\pi_c(n) \geq \pi_f(n-1)$ and no fringe country should have an incentive to join the coalition, e.g. the coalition is externally stable, $\pi_f(n) \geq \pi_c(n+1)$.[10] It is now easy to show that the stable IEA is of size $n^*(\gamma) = I(1/\gamma)$, which is the smallest integer no less than $1/\gamma$. Consider internal stability and consider the non-trivial situation where n members do not pollute because $1 \leq \gamma n$. If after one member left and $1 \leq \gamma(n-1)$ was still true, so the remaining coalition members continue not to pollute, the gain from leaving would be positive: the additional benefit is 1 and the additional damage is γ with $1 > \gamma$ by Assumption 1. Thus, a coalition of n members can only be stable if and only if $1 > \gamma(n-1)$ is true after one member left as the remaining coalition members would switch from $x_c(n) = 0$ to $x_c(n-1) = 1$. Then the additional benefit from pollution of 1 falls short of the additional damage γn, as by assumption $1 \leq \gamma n$ in the initial situation with n members. It is easily checked that such an equilibrium is also externally stable. The total payoff in a stable coalition is given by:

$$\begin{aligned} \Pi(\gamma) &\equiv n^*(\gamma)\pi_c(n^*(\gamma)) + (N - n^*(\gamma))\pi_f(n^*(\gamma)) \\ &= N\pi_0 + (N - n^*(\gamma))(1 - \gamma N) \end{aligned} \quad . \tag{2}$$

Thus, this simple model provides a relationship between the unit damage cost γ and the equilibrium number of coalition members. The equilibrium is a knife-edge equilibrium with $n^*(\gamma)$ countries forming the coalition, which de facto dissolves once a member leaves the coalition as no country would abate anymore. The equilibrium coalition size weakly decreases in the cost-benefit ratio from emissions γ – the larger is γ the smaller is the number of countries in a stable IEA.

2.2 Uncertainty, Risk Aversion and Learning

Now assume that the unit damage cost of global emissions is uncertain and equal for all countries, both ex-ante and ex-post. We denote the value by γ_s in the state of the world s and hence (1) becomes:

$$\pi_{i,s}(x_i, X) \equiv \pi_0 + x_i - \gamma_s X \quad . \tag{3}$$

[10] Without loss of generality, the weak inequality for external stability could be replaced by a strong inequality sign. Our assumption avoids knife-edge cases where a fringe country is indifferent between staying outside and joining a coalition.

Following Kolstad and Ulph (2008) and Boucher and Bramoulle (2010), we assume for simplicity that γ_s can take one of two values: low damage costs, γ_l, with probability p, and high damage costs, γ_h, with probability $1-p$, where $\gamma_l < \gamma_h$ and $0 < p < 1$. We denote by $\bar{\gamma} \equiv p\gamma_l + (1-p)\gamma_h$ the expected value of unit damage costs and by $\bar{\bar{\gamma}} = 0.5(\gamma_l + \gamma_h)$ the median value of damage costs. We define $n_l \equiv I(1/\gamma_l)$, $n_h \equiv I(1/\gamma_h)$, $\bar{n} \equiv I(1/\bar{\gamma})$, $\bar{\bar{n}} \equiv I(1/\bar{\bar{\gamma}})$ and make the following assumptions:

Assumption 2: (i) $1/N < \gamma_l < \gamma_h < 1$; (ii) $2 \le n_h < \bar{\bar{n}} < n_l \le N$.

Assumption 2(i) is just the analogue to Assumption 1(i). Assumption 2(ii) means that uncertainty matters, in the sense that it implies significant differences in the size of the stable IEAs that would arise if we knew for certain which state of the world prevailed.

To allow for risk aversion, we assume that each country has an identical utility function over payoffs: $u(\pi_i), \quad u'(\pi_i) > 0, \quad u''(\pi_i) < 0.$ While ex-ante countries face uncertainty about the true value of unit damage costs, we want to allow for the possibility that countries may learn information during the course of the game which changes the risk they face. We shall follow Kolstad and Ulph (2008) in considering three very simple scenarios of learning, denoted by *m*, $m \in \{NL, PL, FL\}$. With *N*o *L*earning (*m=NL*) countries make their decisions about membership and emissions with uncertainty about the true value of unit damage costs. With *F*ull *L*earning (*m=FL*) countries learn the true value of unit damage costs before they have to take their decisions on membership (stage 1) and emissions (stage 2). With *P*artial *L*earning (*m=PL*) countries learn the true value of damage costs at the end of stage 1, that is after they have made their membership decisions but before they make their emission decisions (stage 2). Thus, in this simple analysis, learning takes the form of revealing perfect information.[11] We will compare the outcomes of the three scenarios of learning in terms of the expected size of IEAs and expected aggregate utility from an ex-ante perspective, i.e. before stage 1.

[11] We use the term "partial learning" in line with the IEA-literature, although this may be misleading; partial learning is usually referred to as delayed but perfect learning, so not partial learning in the sense of Bayesian updating.

3. Analytical Results

In this section, we set out the equilibrium of the IEA model for each of the three models of learning with risk aversion, generalizing the results of Kolstad and Ulph (2008) who assumed risk neutrality. The proofs are provided in the Appendix.

3.1 Full Learning

We start with the benchmark scenario of Full Learning (FL). Players know the realization of the damage parameter γ at the outset of the coalition formation game, i.e. before stage 1. Thus, the results follow directly from what we know from the game with certainty in section 2.1 above.

Proposition 1: Full Learning

If state $s=l,h$ has been revealed before stage 1, then the subsequent membership and emission decisions are as in the model with certainty with damage cost parameter γ_s. In each state $s=l, h$ the size of the stable IEA is $n_s=I(1/\gamma_s)$ and the expected membership is $E(n^{FL})=pn_l+(1–p)n_h$. The expected aggregate utility is given by:

$$E(U^{FL}) = pn_l u(\pi_{c,l}^{FL}) + p(N-n_l)u(\pi_{f,l}^{FL}) + (1-p)n_h u(\pi_{c,h}^{FL}) + (1-p)(N-n_h)u(\pi_{f,h}^{FL})$$

with $\pi_{c,l}^{FL} = \pi_0 - \gamma_l(N-n_l)$, $\pi_{f,l}^{FL} = \pi_0 + 1 - \gamma_l(N-n_l)$, $\pi_{c,h}^{FL} = \pi_0 - \gamma_h(N-n_h)$ *and* $\pi_{f,h}^{FL} = \pi_0 + 1 - \gamma_h(N-n_h)$.

Note that with Full Learning, while the degree of risk aversion does not affect the expected size of the IEA it will affect expected utility. Importantly, the size and utility are computed from an ex-ante perspective to make a comparison with the other models of learning meaningful.

3.2 No Learning

In this section, we address the scenario of No Learning in which players take their membership (stage 1) and emission (stage 2) decisions under uncertainty.[12] We begin by solving for optimal emissions of countries for any number of IEA members *n*. Since the benefit of one unit of emissions exceeds the damage cost in both states of the world, it is

[12] Our analysis of No Learning is similar to the analysis provided by Boucher and Bramoulle (2010).

straightforward to see that fringe countries will always pollute. To solve for the optimal emissions for a coalition member for any *n*, which we denote by $x_c(n)$, we need to introduce some notation. We define:

$$E(u_c(x_c(n),n))) \equiv pu(\pi_{c,l}(x_c(n),n)) + (1-p)u(\pi_{c,h}(x_c(n),n)) \tag{4a}$$

where

$$\pi_{c,s}(x_c(n),n) = \pi_0 - \gamma_s(N-n) + x_c(n)(1-\gamma_s n)\,,\ s = l,h \tag{4b}$$

$E(u_c(x_c(n),n)))$ is the expected utility to an IEA country when there are *n* IEA members who set emissions x_c and all fringe countries set emissions equal to 1. Then:

$$\frac{\partial E(u_c)}{\partial x_c} = p(1-\gamma_l n)u'[\pi_{c,l}(x_c,n)] + (1-p)(1-\gamma_h n)u'[\pi_{c,h}(x_c,n)] \tag{5a}$$

$$\frac{\partial^2 E(u_c)}{\partial (x_c)^2} = p(1-\gamma_l n)^2 u''[\pi_{c,l}(x_c(n),n)] + (1-p)(1-\gamma_h n)^2 u''[\pi_{c,h}(x_c(n),n)] < 0 \tag{5b}$$

From (5a) it is clear that if $\gamma_l n \geq 1$, then $\frac{\partial E(u_c)}{\partial x_c} < 0$ and hence it is optimal for IEA countries to completely abate, while if $\gamma_h n \leq 1$, then $\frac{\partial E(u_c)}{\partial x_c} > 0$ and hence it is optimal for IEA countries to completely pollute. To get tighter bounds on when IEA countries completely pollute or abate, we define $\tilde{n}$ as the largest value of *n* such that:

$$p(1-\gamma_l n)u'[\pi_{c,l}(1,n)] + (1-p)(1-\gamma_h n)u'[\pi_{c,h}(1,n)] \geq 0 \tag{6a}$$

and $\tilde{\tilde{n}}$ as the smallest value of *n* such that:

$$p(1-\gamma_l n)u'[\pi_{c,l}(0,n)] + (1-p)(1-\gamma_h n)u'[\pi_{c,h}(0,n)] \leq 0 \tag{6b}$$

We summarise the results on emissions in the following Lemma:

Lemma 1: Emission Decisions with No Learning

(i) $x_f(n) = 1 \ \forall\, n\,,\ 2 \leq n \leq N$;

(ii) $n_h - 1 \leq \tilde{n} \leq \tilde{\tilde{n}} \leq \bar{n}$

(iii) $p \to 0 \Rightarrow \tilde{n}, \tilde{\tilde{n}}, \bar{n} \to n_h; \quad p \to 1 \Rightarrow \tilde{n}, \tilde{\tilde{n}}, \bar{n} \to n_l$

(iv) $n \le \tilde{n} \Rightarrow x_c(n) = 1; \quad n \ge \tilde{\tilde{n}} \Rightarrow x_c(n) = 0;$

(v) $\tilde{n} \le n < \tilde{\tilde{n}}$

As already noted, for any size of an IEA, *n*, fringe countries always choose the upper limit of emissions. For IEA members, there is a critical range of values for *n*, $[\tilde{n}, \tilde{\tilde{n}}]$, which lies between n_h - 1 and $\bar{n}$ such that IEA members will completely abate if $n \ge \tilde{\tilde{n}}$, and choose the upper limit of emissions for $n \le \tilde{n}$; but if there are values of *n* which lie within the range $]\tilde{n}, \tilde{\tilde{n}}[$, then coalition members choose a level of emissions below the upper limit. Note that, as in Bramoulle and Boucher (2010), with both risk neutrality and risk aversion $x_c(n) = 0$ for $n \ge \bar{n}$ and $x_c(n) = 1$ for $n < \tilde{n} < \bar{n}$. But for $n \le \tilde{n} < \bar{n}$ with risk neutrality $x_c(n) = 1$; while with risk aversion $0 \le x_c(n) < 1$ for $\tilde{n} \le n < \tilde{\tilde{n}}$ and $x_c(n) = 0$ for $\tilde{\tilde{n}} \le n < \bar{n}$. So for $\tilde{n} \le n < \bar{n}$ aggregate emissions are lower with risk aversion than with risk neutrality.

Proposition 2: No Learning

With No Learning, for all parameter values, there exists a stable IEA with membership n^{NL}, which is the same in both states of the world with $n^{NL} \in [\tilde{n}+1, \tilde{\tilde{n}}]$. This is also expected membership, i.e. $E(n^{NL}) = n^{NL} \in [\tilde{n}+1, \tilde{\tilde{n}}]$ which is (weakly) lower than under risk neutrality with $E(n^{NL}) = \bar{n}$ as $\tilde{\tilde{n}} \le \bar{n}$. Emissions of fringe and IEA members are given in Lemma 1. Expected aggregate utility is given by:

$$E(U^{NL}) = pn^{NL}u(\pi_{c,l}^{NL}) + p(N - n^{NL})u(\pi_{f,l}^{NL}) + (1-p)n^{NL}u(\pi_{c,h}^{NL}) + (1-p)(N - n^{NL})u(\pi_{f,h}^{NL})$$

with $\pi_{c,l}^{NL} = \pi_0 - \gamma_l(N - n^{NL}) + (1 - n^{NL}\gamma_l)x^c(n^{NL})$, $\pi_{f,l}^{NL} = \pi_0 - \gamma_l(N - n^{NL}) + 1 - n^{NL}\gamma_l x^c(n^{NL})$, $\pi_{c,h}^{NL} = \pi_0 - \gamma_h(N - n^{NL}) + (1 - n^{NL}\gamma_h)x^c(n^{NL})$ and $\pi_{f,h}^{NL} = \pi_0 - \gamma_h(N - n^{NL}) + 1 - n^{NL}\gamma_h x^c(n^{NL})$.

So the expected equilibrium coalition size is (weakly) smaller under risk aversion than risk neutrality. With uncertainty, countries are unsure about the state of the world. With risk, and hence concave utility, countries shy away from the commitment to be a member in an IEAs members always have lower expected utility than fringe countries. This is in line with the findings in Boucher and Bramoulle (2010).

3.3 Partial Learning

In the scenario of Partial Learning, countries have to make their decision on whether to join an IEA without knowing the true damage cost of emissions, but can make their subsequent emission decisions based on that knowledge. We have argued above that this is the one out of the three scenarios of learning we present which most closely represents the situation the world faces.

It follows that the emission decisions of countries do not depend on risk aversion and so are the same as in Kolstad and Ulph (2008). Since for one unit of emissions the benefit exceeds damage costs in each state of the world, a fringe country will optimally set $x_{f,s}=1$, *s=l,h*; for an IEA member optimal emissions depend on the size of the IEA *n;* so $n \geq n_l \Rightarrow x_{c,s}(n)=0$, $s=l,h$; $n_h \leq n < n_l \Rightarrow x_{c,l}(n)=1$ and $x_{c,h}(n)=0$; $n < n_h \Rightarrow x_{c,s}(n)=1$, $s=l,h$. That is, fringe countries always pollute; if there are at least n_l IEA members, then IEA members always abate; if there are less than n_h IEA members, then IEA members always pollute; otherwise IEA members pollute in the low damage cost state and abate in the high damage cost state.

As in Kolstad and Ulph (2008), for certain values of *p* there may be more than one stable IEA with Partial Learning. In our model a second stable IEA exists iff $\tilde{p} \leq p \leq 1$ where $\tilde{p} \equiv \frac{\chi}{1+\chi}$ with $\chi \equiv \frac{u[\pi_0-(N-n_l)\gamma_h+(1-\gamma_h)]-u[\pi_0-(N-n_l)\gamma_h]}{u[\pi_0-(N-n_l)\gamma_l]-u[\pi_0-N\gamma_l+1]} > 0$. It is straightforward to show that when countries are risk neutral $\tilde{p}=\frac{1-\gamma_h}{n_l\gamma-\gamma_h}$, as in Kolstad and Ulph (2008). However, more generally, we have not been able to determine analytically how $\tilde{p}$ varies with the degree of risk aversion. In section 4.2 we report our findings on this from our simulation results.

Proposition 3: Partial Learning

With Partial Learning, for all parameter values there always exists a stable IEA with $n^{PL_1}=n_h$ *members. All countries pollute in the low damage cost state, while in the high damage cost state coalition members abate and fringe countries pollute. Expected aggregate utility is given by:*

$$E(U^{PL_1}) = pn_h u(\pi_{c,l}^{PL_1}) + p(N-n_h)u(\pi_{f,l}^{PL_1}) + (1-p)n_h u(\pi_{c,h}^{PL_1}) + (1-p)(N-n_h)u(\pi_{f,h}^{PL_1}).$$

with $\pi_{c,l}^{PL_1} = \pi_0 + 1 - \gamma_l N$, $\pi_{f,l}^{PL_1} = \pi_0 + 1 - \gamma_l N$, $\pi_{c,h}^{PL_1} = \pi_0 - \gamma_h(N-n_h)$ *and* $\pi_{f,h}^{PL_1} = \pi_0 + 1 - \gamma_h(N-n_h)$.

If $\tilde{p} \le p \le 1$, *then there is a second stable IEA with* $n^{PL2} = n_l$ *members. In both states of the world, coalition members abate and fringe countries pollute. Expected aggregate utility is given by:*

$$E(U^{PL_2}) = pn_l u(\pi_{c,l}^{PL_2}) + p(N-n_l)u(\pi_{f,l}^{PL_2}) + (1-p)n_l u(\pi_{c,h}^{PL_2}) + (1-p)(N-n_l)u(\pi_{f,h}^{PL_2})$$

with $\pi_{c,l}^{PL_2} = \pi_0 - \gamma_l(N-n_l)$, $\pi_{f,l}^{PL_2} = \pi_0 + 1 - \gamma_l(N-n_l)$, $\pi_{c,h}^{PL_1} = \pi_0 - \gamma_h(N-n_l)$ *and* $\pi_{f,h}^{PL_2} = \pi_0 + 1 - \gamma_h(N-n_l)$.

Since the second equilibrium Pareto-dominates the first equilibrium if it exists, expected membership is either $E(n^{PL_1}) = n_h$ *if* $p < \tilde{p}$ *or* $E(n^{PL_2}) = n_l$ *if* $\tilde{p} \le p \le 1$.

As the degree of risk aversion affects $\tilde{p}$ it has an effect on the likelihood of a second coalition with higher membership n_l being stable. This effect is further explored in section 4.2 where we show that the likelihood of the larger equilibrium decreases with risk aversion.

3.4 Comparison Across the Three Scenarios of Learning

In this sub-section, we investigate what we can say about expected IEA membership, payoffs and expected utility across the four possible equilibria, FL, NL, PL_1 and PL_2.

In terms of expected membership of an IEA it is clear that since $E(n^{PL_1}) = n_h$, this equilibrium has the lowest expected membership while since $E(n^{PL_2}) = n_l$, this equilibrium has the highest expected membership. Note also that:

$$E(n^{FL}) \ge p(\frac{1}{\gamma_l}) + (1-p)(\frac{1}{\gamma_h}) = \frac{p\gamma_h + (1-p)}{\gamma_l \gamma_h}.$$

Moreover it is straightforward to show that:

$$\frac{p\gamma_h + (1-p)\gamma_l}{\gamma_l \gamma_h} > \frac{1}{p\gamma_l + (1-p)\gamma_h} \Leftrightarrow p(1-p)(\gamma_h - \gamma_l)^2 > 0$$

Hence:

$$E(n^{FL}) > \frac{1}{p\gamma_l + (1-p)\gamma_h} \tag{7a}$$

From Proposition 2 and Lemma 1 we have that:

$$E(n^{NL}) \le \tilde{\tilde{n}} \le \bar{n} = I(\frac{1}{p\gamma_l + (1-p)\gamma_h}) \tag{7b}$$

Taken together (7a) and (7b) would suggest that for a wide range of parameters:

$$n^{PL1} \le n^{NL} \le n^{FL} \le n^{PL2} \tag{8a}$$

However, because $E(n^{FL}) = pI(\frac{1}{\gamma_l}) + (1-p)I(\frac{1}{\gamma_h})$, (7a) and (7b) are not sufficient to ensure that $E(n^{NL}) \le E(n^{FL})$, so, as we shall see, there are parameter values for which:

$$E(n^{PL_1}) \le E(n^{FL}) < E(n^{NL}) \le E(n^{PL_2}) \tag{8b}$$

is possible. In terms of payoffs across the four equilibria, it is straightforward to see from Proposition 1, 2 and 3 that:

$$\pi_{c,l}^{PL_2} = \pi_{c,l}^{FL} \ge \pi_{c,l}^{PL_1}; \quad \pi_{f,l}^{PL_2} = \pi_{f,l}^{FL} > \pi_{f,l}^{PL_1}; \quad \pi_{c,h}^{PL_2} > \pi_{c,h}^{FL} = \pi_{c,h}^{PL_1}; \quad \pi_{f,h}^{PL_2} > \pi_{f,h}^{FL} = \pi_{f,h}^{PL_1} \tag{9a}$$

For NL, in the low damage cost state of the world, the highest payoff to coalition members is when x_c=1, which is less than or equal to the payoff to coalition members in PL_2 since $n_l\gamma_l \ge 1$; in the high damage cost state of the world, the highest payoff to coalition members is when x_c = 0, which is less than the payoff to members in PL_2 since $E(n^{NL})\gamma_h < n_l\gamma_h$. So it must be the case that:

$$\pi_{c,l}^{PL_2} \ge \pi_{c,l}^{NL}; \quad \pi_{f,l}^{PL_2} > \pi_{f,l}^{NL}; \quad \pi_{c,h}^{PL_2} > \pi_{c,h}^{NL}; \quad \pi_{f,h}^{PL_2} > \pi_{f,h}^{NL} \tag{9b}$$

Although (9a) and (9b) allow us to rank many of the payoffs across the four possible equilibria for both members and fringe countries in the high and low damage cost states of the world, this is not sufficient to allow us to rank expected aggregate utility at an analytical and general level. The next section reports the simulations we have carried out to compare expected IEA membership and expected welfare across the different models of learning.

4. Results from Simulations

There are three sets of issues we wish to explore using numerical simulations. (i) What is the expected size of the IEA in the case of No Learning, $E(n^{NL})$, in relation to the theoretical limits $\tilde{n}+1$ and $\tilde{\tilde{n}}$ and, more importantly, to the key parameters of our model, n_h and $\bar{n}$ and how does this vary across different degrees of risk aversion? (ii) In the case of Partial Learning what is the critical value of the likelihood of low damage state of the world $\tilde{p}$ such that for $\tilde{p} \le p \le 1$ there is second stable IEA ($n^{PL_2} = n_l$) and how does $\tilde{p}$ vary across different degrees of risk aversion? (iii) How does the expected size of IEA and expected aggregate utility compare across the three different models of learning, Full Learning (FL), No Learning (NL), and Partial Learning (PL_1, PL_2) and how does this comparison depend on the degree of risk aversion?

To address these questions, assume first that each country has a CRRA utility function[13]

$$u(\pi_i) = \frac{A}{(1-\rho)} \pi_i^{1-\rho} \tag{10}$$

where $\rho \ge 0$ measures the degree of relative risk aversion and with A a constant which we will set equal to 1.[14] Moreover, we set $\pi_0 = N$ - 1 in payoff function (1), the smallest value required to ensure non-negative payoffs. We use the risk neutral case (ρ = 0) as a benchmark and then choose 10 values of ρ = 0.1, 0.5, 0.99, 1.5, 2.5, 5.0, 7.0, 10.0, 15.0 and 20.0 to capture what we believe to be a reasonable range of values for country-level risk aversion. For each of these values of risk aversion, we conduct 500,000 Monte Carlo simulations across the values of the remaining key parameters (N, p, γ_l, γ_h). We first choose N as an integer in the range [4,100] and p as a real number in the range [0.001, 0.999]. We then choose γ_l and γ_h to

[13] Meyer and Meyer (2006) note that the CRRA utility function is widely used in empirical studies of risk aversion, and that empirical estimates of ρ vary between 0 and 100. They note that such estimates depend on the variable that enters the utility function, and for the three most commonly used variables – wealth, income and profits – the appropriate empirical estimate increases as one moves from wealth to profits. In our one-period model the relevant variable is income, though there is no distinction between wealth and income. Hence, we have chosen a range of values for ρ at the lower end of the range noted by Meyer and Meyer. See details below. Also note that qualitative conclusions would not change for higher degrees of risk aversion.

[14] The constant A is a multiplicative factor which has no effect on the simulation results presented in this section.

ensure that n, and n_h satisfy Assumption 2 and are evenly distributed between 2 and N. [15] Thus, our simulations basically capture a larger parameter range.

4.1 Results for Size of Stable IEA with No Learning

Recall that from Kolstad and Ulph (2008) the expected size of the stable IEA with No Learning when countries are risk neutral is $E(n^{NL}) = \bar{n}$. In Table 1, we present the results when $\rho > 0$.[16]

Table 1: Expected Size of IEA under no learning for different degrees of risk aversion

1	ρ	0.01	0.50	0.99	1.50	2.50	5.00	7.50	10.00	15.00	20.00
2	% interior Solution	0.06	2.49	4.32	5.88	8.34	12.59	15.30	17.36	19.97	21.72
3	% of cases relating $n^{NL}, \tilde{n}+1, \tilde{\tilde{n}}$										
4	$\tilde{n}+1 = E(n^{NL}) < \tilde{\tilde{n}}$	0.06	2.48	4.30	5.85	8.30	12.50	15.18	17.21	19.80	21.52
5	$\tilde{n}+1 < E(n^{NL}) = \tilde{\tilde{n}}$	0.01	0.32	0.49	0.59	0.71	0.84	0.85	0.85	0.78	0.73
6	$\tilde{n}+1 < E(n^{NL}) < \tilde{\tilde{n}}$	0.00	0.01	0.02	0.03	0.04	0.09	0.12	0.15	0.17	0.20
7	$\tilde{n}+1 = E(n^{NL}) = \tilde{\tilde{n}}$	99.93	97.19	95.18	93.53	90.95	86.57	83.86	81.79	79.25	77.55
8	% of cases relating $n^{NL}, n_h, \bar{n}$:										
9	$n^h = E(n^{NL}) < \bar{n}$	0.04	1.54	2.98	4.32	6.64	11.19	14.68	17.48	21.73	24.98
10	$n_h < E(n^{NL}) = \bar{n}$	83.83	74.59	68.55	63.74	57.08	47.29	41.83	37.63	32.70	29.05
11	$n_h < E(n^{NL}) < \bar{n}$	0.20	7.85	12.53	15.97	20.28	25.45	27.57	28.83	29.65	29.95
12	$n_h = E(n^{NL}) = \bar{n}$	15.93	16.02	15.94	15.97	16.00	16.07	15.92	16.06	15.92	16.02

We know from Proposition 2 that $\tilde{n}+1 \le E(n^{NL}) \le \tilde{\tilde{n}}$, and from Lemma 1($v$), taking account of the previous footnote that if $\tilde{n} < E(n^{NL}) < \tilde{\tilde{n}}$, then $0 < x_c < 1$. Row 2 of Table 1 shows the percentage of cases for which $0 < x_c < 1$ for different values of ρ while rows 4-7 show the percentage of cases for how $E(n^{NL})$ relates to sub-intervals of $[\tilde{n}+1, \tilde{\tilde{n}}]$. It is readily checked that for each value of ρ the percentage in row 2 equals the sum of the percentages in rows 4 and 6, confirming Lemma 1(v). When ρ=0.01, for more than 99.9% of the cases $E(n^{NL}) = \tilde{\tilde{n}}$, i.e. the results are very close to the result under risk neutrality. As risk aversion increases, the

15 Further details are provided in the Appendix.

16 In the proofs of Lemma 1 and Proposition 2 in the Appendix we note that we cannot rule out the theoretical possibility of multiple values of either $\tilde{\tilde{n}}$ or stable IEAs. In the simulations such outcomes occurred in less than 0.01% of parameter combinations, and we have just ignored such cases.

number of cases where $E(n^{NL}) = \tilde{\tilde{n}} \leq \bar{n}$ falls to just over 78% and the number of cases where $E(n^{NL}) = \tilde{n}+1 < \tilde{\tilde{n}}$ rises to just under 22%; in less than 0.2% of the cases do we have $\tilde{n}+1 < E(n^{NL}) < \tilde{\tilde{n}}$.

In terms of the key parameters of the model, n_h and $\bar{n}$, from Lemma 1 we also know that $n_h \leq E(n^{NL}) \leq \bar{n}$ and rows 9-12 in Table 4 show, for each value of ρ, the percentage of sub-cases that can arise. When ρ=0.01, for more than 99.7% of all parameter values, $E(n^{NL}) = \bar{n}$, which is very close to the result with risk neutrality in Kolstad and Ulph (2008). As ρ increases the percentage of cases where $E(n^{NL}) = n_h = \bar{n}$ remains roughly constant at about 16%, reflecting the fact that $p \to 0 \Rightarrow \bar{n} \to n_h$ irrespective of the degree of risk aversion. However for the remaining cases, where $n_h < \bar{n}$, then as risk aversion increases: (i) the percentage of cases where $n_h < E(n^{NL}) = \bar{n}$ falls sharply from 84% to 29%; (ii) the percentage of cases where $n_h < E(n^{NL}) < \bar{n}$ rises sharply from 0.2% to 30% and (iii) the proportion of cases where $n_h = E(n^{NL}) < \bar{n}$ rises steadily from 0.04% to 25%. Hence, with No Learning, increasing risk aversion drives down the number of countries joining an IEA as discussed in our theoretical analysis in subsection 3.2.

4.2 Second stable IEA with Partial Learning

We showed in section 3.3 that with Partial Learning there always exists a stable IEA with n_h members who abate in the high damage cost state and pollute in low damage cost state, but there is a critical value of p, which we defined as $\tilde{p}$, such that for $\tilde{p} \leq p \leq 1$ there exists a second stable IEA where members abate in both states of the world. We also said that we had been unable to prove analytically how $\tilde{p}$ varies with the degree of risk aversion. In rows 2 and 3 of Table 2, we show, for each value of risk aversion between 0 and 20, the average value of $\tilde{p}$ and the percentage of such cases for which $\tilde{p} \leq p \leq 1$ occurs, respectively. From row 2 we see that as the degree of risk aversion increases, the average value of $\tilde{p}$ rises from 0.9437 to 0.9685, while from row 3 the percentage of simulations for which $\tilde{p} \leq p \leq 1$ falls from 5.66% to 3.14%. So increasing risk aversion reduces the likelihood that with Partial Learning there exists a second stable IEA with higher membership.

Table 2: Expected membership and aggregate utility for three scenarios of learning

1	ρ	0.00	0.10	0.50	0.99	1.50	2.50	5.00	7.50	10.0	15.0	20.0
2	Average $\tilde{p}$	.9437	.9442	.9472	.9508	.9539	.9586	.9645	.9671	.9683	.9689	.9685
3	% Cases where $p \geq \tilde{p}$	5.66	5.56	5.27	4.87	4.59	4.12	3.52	3.29	3.17	3.08	3.14
4	% Cases for membership											
5	$E(n^{PL})$ highest	5.66	5.56	5.27	4.87	4.59	4.12	3.52	3.29	3.17	3.08	3.14
6	$E(n^{NL})$ highest	15.07	14.99	14.60	14.21	13.94	13.33	12.33	11.55	11.01	10.20	9.44
7	$E(n^{FL})$ highest	79.27	79.45	80.13	80.92	81.47	82.55	84.15	85.16	85.82	86.72	87.42
8	% Cases for aggregate utility:											
9	$E(U^{PL})$ highest	5.66	5.56	5.27	4.87	4.59	4.12	3.52	3.29	3.17	3.08	3.14
10	$E(U^{NL})$ highest	72.90	72.43	70.42	68.26	66.05	62.42	54.95	49.32	44.79	38.25	33.43
11	$E(U^{FL})$ highest	21.44	22.01	24.33	26.87	29.36	33.46	41.53	47.39	52.04	58.67	63.43

4.3 Comparison Across the Three Scenarios of Learning

In the remaining part of Table 2, we present for each value of ρ between 0 and 20 the percentage of simulations for which we have particular rankings for expected membership and expected aggregate utility across our three models of learning. In the case of PL, if $\tilde{p} \leq p \leq 1$ and there are two stable IEAs, then we chose the second Pareto-superior equilibrium.

The first result we note is that for both expected membership and expected aggregate utility, the percentage of simulations where PL performs best is equal to the percentage of simulations for which $\tilde{p} \leq p \leq 1$ as discussed in section 4.2. This shows that Partial Learning is the preferred model of learning only in those cases where $\tilde{p} \leq p \leq 1$, i.e. if the large coalition equilibrium emerges. As we have already noted, the percentage of such cases declines from just under 5.66% when ρ=0 to just 3.14% when ρ = 20.

Focusing now on just FL and NL, as ρ increases from 0 to 20, the percentage of simulations where FL has the highest expected membership rises from 79.27% to 87.42% while the percentage of simulations where NL has the highest expected membership decreases from 15.07 % to 9.44%. This is consistent with the result presented in Table 1, namely that as risk aversion increases the size of the IEA with NL decreases from $\bar{n}$ towards n_h. More importantly, as risk aversion increases, the percentage of simulations where expected aggregate utility is strictly highest under NL decreases from just under 73% to just over 33%,

while the percentage of simulations where FL has highest expected aggregate utility increases from just over 21% to just over 63%.

The implication of these results is that as countries become more risk averse, then for an increasing percentage of parameter values the form of learning which leads to the highest aggregate utility is Full Learning. The percentage of parameter values favouring No Learning decreases, and for high enough values of risk aversion the percentage of cases for which Full Learning is the preferred scenario is higher than No Learning. This is also true for the expected utility of an individual country from an ex-ante perspective, which, for symmetric players, is simply the aggregate utility divided by the total number of countries. Thus, could governments choose the scenario of learning endogenously in stage zero, they preferred full learning for high levels of risk aversion. So countries would be better off leaving the decision to form an IEA and set their emissions until they have Full Information about the risks of climate change. The relatively small percentage of parameter values for which the preferred model of forming an IEA is Partial Learning, i.e. deciding whether to join an IEA before getting better information about the risks of climate change, but allowing emissions to be set after better information is available, also declines as countries become more risk averse. Thus, the pessimistic conclusion about the role of learning in a strategic context derived in previous papers is qualified, in particular if the degree of risk aversion is sufficiently large.

5. Summary and Conclusions

This paper bridges two strands of literature on the formation of IEAs under uncertainty by addressing the combined roles of learning and risk aversion. This approach allowed us to explore the impact of learning for any given level of risk aversion as well as the impact of changing risk aversion under various scenarios of learning.

We generalized the model of Kolstad and Ulph (2008) who showed that with risk neutrality the possibility of learning more information about environmental damage costs generally had rather pessimistic implications for the success of the formation of IEAs. Except for a relatively small set of parameter values for which partial learning would select a high IEA membership, learning resulted in lower or equal expected membership for partial learning and lower expected aggregate and individual payoff for partial and full learning, compared to no learning. Moreover, this parameter range required that the probability of low damage cost is very high a rather uninteresting parameter constellation in the context of climate change. Hence, in a strategic context, learning reduces expected aggregate and individual payoffs for a

wide range of parameter values. Across the different models of learning, they showed that for a large set of parameter values the scenario of learning *which* yielded highest expected aggregate and individual payoff was No Learning, which would suggest that countries are better off forming an IEA rather than waiting for better information.

In this paper, we have allowed countries to be risk averse using an expected utility approach which maps payoffs into utility. We first derived the theoretical results for each of our three scenarios of learning with risk aversion, confirming the main findings of Boucher and Bramoulle (2010) for the No Learning case. For No learning, risk leads to smaller stable coalitions and higher global emissions. In terms of equilibrium coalitions and global emissions, we showed that Full Learning remains unaffected by risk and changes for Partial Learning are small. However, even with special functional forms for the underlying utility functions there was limited scope for deriving analytical comparisons across our three scenarios of learning, primarily because welfare effects could differ for signatory and non-signatory countries. Our simulation results showed that contrary to the finding with risk neutrality, when countries become significantly risk averse, the set of parameter values for which countries are better off with No Learning compared to Full Learning shrinks significantly and those cases for which this is reversed increases accordingly. This may explain why it has taken so long for a proper climate agreement to be reached – countries are risk averse and waited till they had much better information about the risks of climate change.

In terms of future research, it would be desirable to use a model with asymmetric countries, though it is unlikely to be possible to derive analytical results; so it may be more useful to introduce different models of learning into Integrated Assessment Models of climate change. It would also be interesting to endogenise the process of learning by allowing countries to invest in research in order to obtain better information.

References

Arrow, K. and A. Fisher (1974), Environmental preservation, uncertainty and irreversibility. *Quarterly Journal of Economics*, **88**: 312-319.

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

Boucher, V. and Y. Bramoullé (2010), Providing global public goods under uncertainty. *Journal of Public Economics*, **94**: 591-603.

Bramoullé, Y. and N. Treich (2009), Can uncertainty alleviate the commons problem? *Journal of the European Economic Association*, **7(5)**: 1042-1067.

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

Dellink, R. and M. Finus (2012), Uncertainty and climate treaties: does ignorance pay? *Resource and Energy Economics*, **34:** 565-584.

Endres, A. and C. Ohl (2003), International environmental cooperation with risk aversion. *International Journal of Sustainable Development*, **6**: 378-392.

Epstein, L. (1980), Decision-making and the temporal resolution of uncertainty. *International Economic Review*, **21**: 269-284.

Finus, M. and A. Caparrós (2015), *Game Theory and International Environmental Cooperation: Essential Readings*. Edward Elgar, Cheltenham, UK.

Finus, M. and P. Pintassilgo (2013), The role of uncertainty and learning for the success of international climate agreements. *Journal of Public Economics*, 103: 29-43.

Gollier, C., B. Jullien and N. Treich (2000), Scientific progress and irreversibility: an economic interpretation of the 'Precautionary Principle'. *Journal of Public Economics*, **75**: 229-253.

Hong, F. and L. Karp (2014), International environmental agreements with exogenous and endogenous Risk. *Journal of the Association of Environmental and Resource Economics*, **1** 365 – 394.

Kolstad, C. (1996a), Fundamental irreversibilities in stock externalities. *Journal of Public Economics*, **60**: 221-233.

Kolstad, C. (1996b), Learning and stock effects in environmental regulations: the case of greenhouse gas emissions. *Journal of Environmental Economics and Management*, **31**: 1-18.

Kolstad, C. (2007), Systematic uncertainty in self-enforcing international environmental agreements. *Journal of Environmental Economics and Management*, **53**: 68-79.

Kolstad, C. and A. Ulph (2008), Learning and international environmental agreements. *Climatic Change*, **89**: 125-141.

Kolstad, C. and A. Ulph (2011), Uncertainty, learning and heterogeneity in international environmental agreements. *Environmental and Resource Economics,* **50**, 389-403 .

Markowitz, H. (1952), Portfolio selection. *Journal of Finance*, **7**: 77-91.

Meyer, J. (1987), Two-moment decision models and expected utility maximization. *American Economic Review*, **77**: 421-430.

Meyer, D. and J. Meyer (2006), Measuring risk aversion. *Foundations and Trends in Microeconomics*, **2**: 107-203.

Na, S.-L. and H.S. Shin (1998), International environmental agreements under uncertainty. *Oxford Economic Papers*, **50**: 173-185.

Narain, U., A. Fisher and M. Hanemann (2007), The irreversibility effect in environmental decisionmaking. *Environmental and Resource Economics,* **38**: 391-405.

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

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

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

Ulph, A. (2004), Stable international environmental agreements with a stock pollutant, uncertainty and learning. *Journal of Risk and Uncertainty*, **29**: 53-73.

Ulph, A. and D. Maddison (1997), Uncertainty, learning and international environmental policy coordination. *Environmental and Resource Economics*, **9**: 451-466.

Ulph, A. and D. Ulph (1996), Who gains from learning about global warming? In: van Ierland, E. and K. Gorka (eds.), *The Economics of Atmospheric Pollution*. Springer, Heidelberg, ch. 3, 31-62.

Ulph, A. and D. Ulph (1997), Global warming, irreversibility and learning. *Economic Journal*, **107**: 636-650.

Weitzman, M. (2009), On modeling and interpreting the economics of catastrophic climate change. *Review of Economics and Statistics*, **91**: 1-19.

Appendix: Proofs of Results

Proposition 1: Full Learning

Since the true state of the world is revealed before countries decide whether to join an IEA, the results for stable IEAs, probabilities and payoffs in the four states of the world shown in Proposition 1 follow immediately from Kolstad and Ulph (2008). The difference from Kolstad and Ulph is that we now calculate aggregate expected utility using the expected utility approach which captures attitudes to risk, rather than expected payoff. This completes the proof of Proposition 1.

Lemma 1: Emission Decisions with No Learning.

Fringe Country Output Decision

In stage 2, fringe country i takes as given the output of all other countries, X_{-i}, and chooses x_i to maximise $pu[\pi_0 + x_i - \gamma_l(x_i + X_{-i})] + (1-p)u[\pi_0 + x_i - \gamma_h(x_i + X_{-i})]$. Since $\gamma_l < \gamma_h < 1$, $x^f(n) = 1 \ \forall \ n$, $2 \le n \le N$. This proves result (i).

Member Country Emission Decision

(i) Properties of Payoff Function

From (4b) we obtain the following:

$$0 < \pi_{c,h}(x_c, n) < \pi_{c,l}(x_c, n) \ \forall \ n \ge 1, 0 \le x_c \le 1; \tag{A1a}$$

$$\frac{\partial \pi_{c,s}}{\partial x_c} \ge 0 \Leftrightarrow n \le \frac{1}{\gamma_s}; \tag{A1b}$$

$$\pi_{c,s}(0,n) = \pi_0 - N\gamma_s + n\gamma_s; \qquad \pi_{c,s}(1,n) = \pi_0 - N\gamma_s + 1 \equiv \pi_{c,s}(1) \quad \forall n \tag{A1c}$$

and hence

$$\pi_{c,h}(1) < \pi_{c,h}(0,n) < \pi_{c,l}(0,n) < \pi_{c,l}(1) \ \text{ if } \frac{1}{\gamma_h} < n < \frac{1}{\gamma_l} \tag{A1d}$$

As stated in section 3.2, for a given n, each coalition member chooses $x_c(n)$ to maximise

$E(u_c(x_c(n), n))) \equiv pu(\pi_{c,l}(x_c(n), n)) + (1-p)u(\pi_{c,h}(x_c(n), n))$ which leads to the following first and second order condition:

$$\frac{\partial E(u_c)}{\partial x_c} = p(1-\gamma_l n)u'[\pi_{c,l}(x_c,n)] + (1-p)(1-\gamma_h n)u'[\pi_{c,h}(x_c,n)] \quad \text{(A2a)}$$

$$\frac{\partial^2 E(u_c)}{\partial (x_c)^2} = p(1-\gamma_l n)^2 u''[\pi_{c,l}(x_c(n),n)] + (1-p)(1-\gamma_h n)^2 u''[\pi_{c,h}(x_c(n),n)] < 0 \quad \text{(A2b)}$$

Boundary Values for $x_c(n)$

From (A2a):

$$n \geq \frac{1}{\gamma_l} > \frac{1}{\gamma_h} \Rightarrow \frac{\partial E(u_c)}{\partial x_c} < 0 \; \forall \; x_c \Rightarrow x_c(n) = 0$$
$$n \leq \frac{1}{\gamma_h} < \frac{1}{\gamma_l} \Rightarrow \frac{\partial E(u_c)}{\partial x_c} > 0 \; \forall \; x_c \Rightarrow x_c(n) = 1$$

We want to see if we can get tighter bounds for n that guarantees when $x_c(n) = 1$ and $x_c(n) = 0$. So we now focus on the range $n_h - 1 < \frac{1}{\gamma_h} < n < \frac{1}{\gamma_l} \leq n_l$. From (A1b), in this range

$$\frac{\partial \pi_{c,l}}{\partial x_c} > 0; \quad \frac{\partial \pi_{c,h}}{\partial x_c} < 0$$

To make progress, we treat n as if it was a real value, z. To save notation define:

$\tilde{\theta}_l \equiv u'[\pi_{c,l}(1)]$, $\tilde{\theta}_h \equiv u'[\pi_{c,h}(1)]$, $\tilde{\tilde{\theta}}_l(z) \equiv u'[\pi_{c,l}(0,z)]$ and $\tilde{\tilde{\theta}}_h(z) \equiv u'[\pi_{c,h}(0,z)]$ where, from (A1d):

$$\tilde{\theta}_l < \tilde{\tilde{\theta}}_l(z) < \tilde{\tilde{\theta}}_h(z) < \tilde{\theta}_h \quad \text{(A3)}$$

We first define $\tilde{z}$ as the unique value of z such that:

$$\frac{\partial E(u_c(1,\tilde{z}))}{\partial x_c} = p(1-\gamma_l \tilde{z})\tilde{\theta}_l + (1-p)(1-\gamma_h \tilde{z})\tilde{\theta}_h = 0 \Rightarrow \tilde{z} = \frac{p\tilde{\theta}_l + (1-p)\tilde{\theta}_h}{\gamma_l p\tilde{\theta}_l + \gamma_h(1-p)\tilde{\theta}_h}. \quad \text{(A4)}$$

From (A4) we get:

$\frac{1}{\gamma_h} < \tilde{z} < \frac{1}{\gamma_l}$ and $\frac{\partial E(u_c(1,\tilde{z}))}{\partial x_c} > 0 \quad \forall \; z < \tilde{z}$. Thus, $x_c(z) = 1 \; \forall \; z \leq \tilde{z}$.

We now define $\tilde{\tilde{z}}$ such that:

$$\frac{\partial E(u_c(0,\tilde{\tilde{z}}))}{\partial x_c} = p(1-\gamma_l\tilde{\tilde{z}})\tilde{\tilde{\theta}}_l(\tilde{\tilde{z}}) + (1-p)(1-\gamma_h\tilde{\tilde{z}})\tilde{\tilde{\theta}}_h(\tilde{\tilde{z}}) = 0 \Rightarrow \tilde{\tilde{z}} = \frac{p\tilde{\tilde{\theta}}_l(\tilde{\tilde{z}}) + (1-p)\tilde{\tilde{\theta}}_h(\tilde{\tilde{z}})}{p\gamma_l\tilde{\tilde{\theta}}_l(\tilde{\tilde{z}}) + (1-p)\gamma_h\tilde{\tilde{\theta}}_h(\tilde{\tilde{z}})} \quad \text{(A5)}$$

and again $\frac{1}{\gamma_h} < \tilde{\tilde{z}} < \frac{1}{\gamma_l}$. Note, we have not been able to prove that there is a unique value of $\tilde{\tilde{z}}$ which solves (A5). We will discuss the implications shortly, but the next steps apply to any $\tilde{\tilde{z}}$ which solves (A5). We first show that:

$$\begin{aligned}\tilde{\tilde{z}} - \tilde{z} &= [p\tilde{\tilde{\theta}}_l(\tilde{\tilde{z}}) + (1-p)\tilde{\tilde{\theta}}_h(\tilde{\tilde{z}})][p\gamma_l\tilde{\theta}_l + (1-p)\gamma_h\tilde{\theta}_h] \\ -[p\tilde{\theta}_l + (1-p)\tilde{\theta}_h][p\gamma_l\tilde{\tilde{\theta}}_l(\tilde{\tilde{z}}) + (1-p)\gamma_h\tilde{\tilde{\theta}}_h(\tilde{\tilde{z}})] &= p(1-p)(\gamma_h-\gamma_l)\tilde{\theta}_l\tilde{\tilde{\theta}}_l(\tilde{\tilde{z}})[\frac{\tilde{\theta}_h}{\tilde{\theta}_l} - \frac{\tilde{\tilde{\theta}}_h(\tilde{\tilde{z}})}{\tilde{\tilde{\theta}}_l(\tilde{\tilde{z}})}]\end{aligned} \quad \text{(A6)}$$

Using (A3), it can be shown that:

$$sign\left(\tilde{\tilde{z}} - \tilde{z}\right) = sign\left\{ p(1-p)(\gamma_h-\gamma_l)\tilde{\theta}_l\tilde{\tilde{\theta}}_l(\tilde{\tilde{z}})[\frac{\tilde{\theta}_h}{\tilde{\theta}_l} - \frac{\tilde{\tilde{\theta}}_h(\tilde{\tilde{z}})}{\tilde{\tilde{\theta}}_l(\tilde{\tilde{z}})}] \right\} \geq 0.$$ So any $\tilde{\tilde{z}}$ which solves (A5) must be at least as large as $\tilde{z}$.

Next we show that $\tilde{\tilde{z}} \leq \frac{1}{\bar{\gamma}}$. Define $\xi \equiv \tilde{\tilde{\theta}}_h(\tilde{\tilde{z}}) / \tilde{\tilde{\theta}}_l(\tilde{\tilde{z}}) \geq 1$. Then:

$$\begin{aligned}&\tilde{\tilde{z}} \leq 1/\bar{\gamma} \Leftrightarrow \frac{p+(1-p)\xi}{p\gamma_l + (1-p)\gamma_h\xi} \leq \frac{p+(1-p)}{p\gamma_l + (1-p)\gamma_h} \\ &\Leftrightarrow p(1-p)(\gamma_h + \xi\gamma_l) \leq p(1-p)(\xi\gamma_h + \gamma_l) \Leftrightarrow (\xi-1)(\gamma_h-\gamma_l) \geq 0\end{aligned} \quad \text{(A7)}$$

So any $\tilde{\tilde{z}}$ which solves (A5) must be no greater than $1/\bar{\gamma}$.

As $\tilde{n} = \begin{cases} \tilde{z} & if \quad \tilde{z} \text{ is an integer} \\ I(\tilde{z}) - 1 & if \quad \tilde{z} \text{ is not an integer} \end{cases}$ and $\tilde{\tilde{n}} = \begin{cases} \tilde{\tilde{z}} & if \quad \tilde{\tilde{z}} \text{ is an integer} \\ I(\tilde{\tilde{z}}) + 1 & if \ \tilde{\tilde{z}} \text{ is not an integer} \end{cases}$

then $n_h - 1 \leq \tilde{n} \leq \tilde{\tilde{n}} \leq \bar{n}$ - result (ii); $n \leq \tilde{n} \Rightarrow x_c(n) = 1; \quad n \geq \tilde{\tilde{n}} \Rightarrow x_c(n) = 0$ - result (iv).

Finally, from (A4) and (A5) it is straightforward to see that:

$p \to 0 \Rightarrow \tilde{n}, \tilde{\tilde{n}}, \bar{n} \to n_h; \quad p \to 1 \Rightarrow \tilde{n}, \tilde{\tilde{n}}, \bar{n} \to n_l$ - result (iii).

We have not been able to prove analytically that there is a unique value of $\tilde{\tilde{z}}$, so in what follows we will treat $\tilde{\tilde{z}}$ and hence $\tilde{\tilde{n}}$ as the *largest* such value.

(ii) Interior values for $x^c(n)$

It follows from the definitions of $\tilde{n}$ and $\tilde{\tilde{n}}$ that $n > \tilde{n} \Rightarrow x_c(n) < 1$ and $n < \tilde{\tilde{n}} \Rightarrow x_c(n) \geq 0$. Hence, $\tilde{n} < n < \tilde{\tilde{n}} \Rightarrow 0 \leq x_c(n) < 1$ - result (v). This completes the proof of Lemma 1.

Proposition 2: No Learning

From Lemma 1 we know the values of $x_f(n)$ and $x_c(n)$ for all n. We now solve Stage 1. Define:

$E(u_f(n)) = pu(\pi_0 + 1 - \gamma_l[N - n + nx_c(n)]) + (1-p)u(\pi_0 + 1 - \gamma_h[N - n + nx_c(n)])$,

$E(u_c(n)) = pu(\pi_0 + x_c(n) - \gamma_l[N - n + nx^c(n)]) + (1-p)u(\pi_0 + x_c(n) - \gamma_h[N - n + nx_c(n)])$ and

$\Delta(n) = E(u_c(n)) - E(u_f(n-1))$, noting that $n^{NL} = E(n^{NL})$ is a stable IEA iff $\Delta(n^{NL}) \geq 0$ and $\Delta(n^{NL} + 1) < 0$.

To save notation, define $\pi_s = \pi_0 - \gamma_s N \quad s = l, h$.

(i) $n \geq \tilde{\tilde{n}} + 1$

$$\Delta(n) = pu(\pi_l + \gamma_l n) + (1-p)u(\pi_h + \gamma_h n) - pu(\pi_l + \gamma_l n + (1-\gamma_l)) - (1-p)u(\pi_h + \gamma_h n + (1-\gamma_h)) < 0$$

So no such n could be stable.

(ii) $n \leq \tilde{n}$

All countries set $x = 1$, so $\Delta(n) = 0$. We exclude these trivial cases.

(iii) $\tilde{n} + 2 \leq n < \tilde{\tilde{n}}$

$$\begin{aligned}\Delta(n) = {} & p[\pi_l + n\gamma_l + x_c(n)(1-\gamma_l) - \gamma_l(n-1)x_c(n)] \\ & + (1-p)[\pi_h + n\gamma_h + x_c(n)(1-\gamma_h) - \gamma_h(n-1)x_c(n)] \\ & - p[\pi_l + n\gamma_l + (1-\gamma_l) - \gamma_l(n-1)x_c(n-1)] \\ & - (1-p)[\pi_h + n\gamma_h + (1-\gamma_h) - \gamma_h(n-1)x_c(n-1)]\end{aligned}$$

So, *if* $x_c(n) \geq x_c(n-1)$, then $\Delta(n)<0$ and so n could not be a stable IEA. If $x_c(n) < x_c(n-1)$, then the sign of $\Delta(n)$ depends on the precise values of $x_c(n)$ and $x_c(n-1)$.

(iv) $n = \tilde{n} + 1$

From (ii) we know that $x_c(\tilde{n}) = x_f(\tilde{n}) = 1 \Rightarrow E(u_c(\tilde{n})) = E(u_f(\tilde{n}))$. So when $n = \tilde{n} + 1$, IEA members could have set $x_c(\tilde{n}+1) = 1$, but, by definition of $\tilde{n}$, they did not, so $E(u_c(\tilde{n}+1)) > E(u_c(\tilde{n})) = E(u_f(\tilde{n})) \Rightarrow \Delta(\tilde{n}+1) > 0$.

To complete the argument, successively increase *n* from $\tilde{n}+1$ until $\Delta(n) < 0$, in which case $n^{NL} = n\text{-}1$ is a stable IEA. There must exist such a stable IEA since we know from (iv) that $\Delta(\tilde{n}+1) > 0$ and from (i) that $\Delta(\tilde{\tilde{n}}+1) < 0$. We cannot rule out the possibility that there is more than one stable IEA. This completes the proof of Proposition 2.

Proposition 3: Partial Learning

Because emission decisions in stages 2 are taken under certainty, the results are the same as in Kolstad and Ulph (2008), namely:

Stage 2: (a) $x_{f,s}(n) = 1 \quad s = l, h \quad \forall n$

(b) $n \geq n_l \Rightarrow x_{c,s}(n) = 0 \quad s = l, h$

(c) $n_h \leq n < n_l \Rightarrow x_{c,h} = 0, x_{c,l} = 1$

(d) $n < n_h \Rightarrow x_{c,s}(n) = 1 \quad s = l, h$

So using the notation $\pi_s = \pi_0 - \gamma_s N$, $s = l, h$, introduced in the proof of Proposition 2, the payoffs are:

(i) $n \geq n_l$

$$E(u_f(n)) = pu[\pi_l + n\gamma_l + 1] + (1-p)u[\pi_h + n\gamma_h + 1]$$
$$E(u_c(n)) = pu[\pi_l + n\gamma_l] + (1-p)u[\pi_h + n\gamma_h]$$

(ii) $n_h \leq n < n_l$

$$E(u_f(n)) = pu[\pi_l + 1] + (1-p)u[\pi_h + n\gamma_h + 1]$$
$$E(u_c(n)) = pu[\pi_l + 1] + (1-p)u[\pi_h + n\gamma_h]$$

(iii) $n < n_h$

$$E(u_f(n)) = E(u_c(n)) = pu[\pi_l + 1] + (1-p)u[\pi_h + 1]$$

Stage 1:

(i) $n \geq n_l + 1$

$$\Delta(n) = pu(\pi_l + n\gamma_l) + (1-p)u(\pi_h + n\gamma_h) - pu(\pi_l + n\gamma_l + (1-\gamma_l)) \\ -(1-p)u(\pi_h + n\gamma_h + (1-\gamma_h)) < 0$$

So no *n* in this range can be a stable IEA.

(ii) $n = n_l$

$$\Delta(n_l) = pu(\pi_l + n_l\gamma_l) + (1-p)u(\pi_h + n_l\gamma_h) - pu(\pi_l + 1])(1-p)u(\pi_h + n_l\gamma_h + (1-\gamma_h))$$

As $n_l\gamma_l \geq 1$ so:

$$\Delta(n_l) \geq 0 \Leftrightarrow \frac{p}{(1-p)} \geq \frac{u(\pi_h + n_l\gamma_h + (1-\gamma_h)) - u(\pi_h + n_l\gamma_h)}{u(\pi_l + n_l\gamma_l) - u(\pi_l + 1)} \equiv \chi > 0$$

i.e. $\Delta(n_l) \geq 0 \Leftrightarrow p \geq \frac{\chi}{1+\chi} \equiv \tilde{p} > 0$

Since, from (i), $\Delta(n_l + 1) < 0$, n_l is stable iff $p \geq \tilde{p}$.

(iii) $n_h + 1 \leq n < n_l$

$$\Delta(n) = pu(\pi_l + 1) + (1-p)u(\pi_h + n_h\gamma_h) - pu(\pi_l + 1) - (1-p)u(\pi_h + n\gamma_h + (1-\gamma_h)) < 0$$

So no *n* in this range can be stable.

(iv) $n = n_h$

$$\Delta(n_h) = pu(\pi_l + 1) + (1-p)u(\pi_h + n_h\gamma_h) - pu(\pi_l + 1) - (1-p)u(\pi_h + 1)$$

Since $n_h\gamma_h \geq 1$, $\Delta(n_h) \geq 0$ and since, from (iii) $\Delta(n_h + 1) < 0$, n_h is always a stable IEA.

(v) $n < n_h$

$\Delta(n) = 0$; as in Proposition 2, we ignore these cases.

So there always exists a stable IEA with $n^{PL} = n_h$ and if $p \geq \tilde{p}$ there is a second stable IEA with $n^{PL} = n_l$. This completes the proof of Proposition 3.

Further Details about Simulations in Section 4

We do four steps: (i) we choose randomly the inverse of the median, $1/\bar{\bar{\gamma}}$ lying in the range [2, *N –1*] and calculate $\bar{\bar{n}} = I(1/\bar{\bar{\gamma}})$; (ii) we select the smaller of the two ranges $]1, \bar{\bar{n}} - 1[$ ($]\bar{\bar{n}}, N[$, respectively) and choose randomly $1/\gamma_h$ ($1/\gamma_l$, respectively} to lie in that range, and hence set $n_h = I(1/\gamma_h)$ ($n_l = I(1/\gamma_l)$ respectively); (iii) we next calculate n_l (n_h, respectively) to satisfy $n_l + n_h = 2\bar{\bar{n}}$; (iv) finally, we choose randomly $1/\gamma_l$ ($1/\gamma_h$, respectively) to lie in the range $]n_l - 1, n_l[$ ($]n_h - 1, n_h[$, respectively). The result of these calculations is that γ_l lies in the range [0.0100, 0.3333], with an average value over the simulations of 0.0571, and γ_h lies in the range [0.0102,0.9999] with an average value of 0.1662.