The Effectiveness of International Environmental Agreements*

Jürg Vollenweider†

March 20, 2012

Abstract

Many argue that International Environmental Agreements (IEAs) can alter states' cost-benefit analyses by providing crucial information about the costs of environmental degradation. Thereby, IEAs may help to effectively curb environmental pollution. However, previous attempts to empirically measure institutional effectiveness have largely failed to provide credible estimates because they have missed to produce convincing counterfactuals. This study empirically estimates the effectiveness of one prominent example of an international environmental institution, the Long Range Transboundary Air Pollution agreement (LRTAP). It sets forth a transparent identification strategy in light of latest advancements in the causal inference literature. The study presents evidence for the non-effectiveness of the LRTAP in changing member states' behavior with regard to reducing anthropogenic emissions of two substances, NO_x and SO_2. Additionally, by identifying a regression difference in difference (DID) design, it theoretically derives a convincing counterfactual and establishes a methodological tool-kit to draw causal inferences about the effectiveness of international environmental institutions in general.

Keywords: Institutional Effectiveness, Causal Inference, Difference-in-Differences, International Environmental Agreements, Air Pollution

*This work was supported by 'ETH Independent Investigators' Research Award' (ETHIIRA) grant ETH-14 09-3. This article benefited from comments of Michael M. Bechtel, Tobias Böhmelt, Beatrice Brunner, the audience at the 2011 annual meeting of the International Political Economy Society, and members of the international political economy research group at ETH Zurich. The usual disclaimer applies.

†ETH Zurich, Center for International and Comparative Studies (CIS), IFW C 43.1, Haldeneggsteig 4, 8092 Zurich; jvollenweider@ir.gess.ethz.ch

Introduction

Since the 1970's International Environmental Agreements (IEAs) have gained prominence as means to tackle various international problems of negative externalities that harm the environment. In the three decades following the 1970s the number of international environmental institutions has grown by more than 170 % to 464 agreements. Countries have created environmental agreements with a wide range of targets, from reducing oil or air pollution to protecting polar bears (see e.g. Haas *et al.*, 1993; Miles *et al.*, 2002; Helm and Sprinz, 2000; Hovi *et al.*, 2003; Breitmeier *et al.*, 2006). Also, this proliferation reflects a growing public concern with regard to environmental degradation. Most prominently, climate change, and what many consider its most important contributer, anthropogenic air pollution, lies at the heart of the public and scientific debate.

The global scientific community is largely in agreement that air pollution constitutes a serious threat to human health and ecosystems in most parts of the world. Since the 1960s, scientists have gradually established that a substantial amount of air pollution travels over long distances and also crosses international boundaries (see e.g. Odén, 1968; Andreaevc and Crutzen, 1997). Hence, it does not come as a surprise that more than 40 international environmental institutions exist to reduce harmful air pollution.

Given the increasing use and importance of IEAs, a debate evolved about their effectiveness in altering states' behavior. The question if international environmental institutions really succeed in effectively altering states' behavior is important in at least three ways. First, if states are serious about environmental protection it is crucial to have knowledge about the effects of IEAs to improve the design of such institutions. States can only allocate resources efficiently if they can base their decisions on credible effectiveness measures. Second, electoral accountability suggests that citizens' electoral choices at least partly reflect information about their government's performance in protecting the environment. Finally, although many offer policy recommendations that advise the use of certain types of environmental institutions to address environmental problems, this advice is largely based on theoretical contributions and rests on weak empirical foundations.[1] For example, game theorists adopt predominantly a critical view on the effectiveness of IEAs. Because states act out of self-interest, the level of environmental protection does not exceed what we would expect in the absence of this institution (Murdoch and Sandler, 1997; Maler and Zeeuw, 1998; Barrett, 2003). Others, however, argue that states'

[1] Downs (2000), Barrett (2003), and Ringquist and Kostadinova (2005) provide excellent discussions of the existing theoretical arguments.

preferences can change through participation in an environmental agreement (Downs, 2000). Ultimately, the test of these arguments requires convincing measures of institutional effectiveness.

Although researchers have already invested some effort in empirically exploring the question whether international environmental agreements are effective instruments for protecting the environment, the scientific debate seems far from moving toward closure (Bernauer, 1995; Victor *et al.*, 1998; Young, 1999; Downs, 2000; Helm and Sprinz, 2000; Young, 2001; Miles *et al.*, 2002; Mitchell, 2002; Hovi *et al.*, 2003; Young, 2003; Mitchell, 2006; Böhmelt and Pilster, 2010; Breitmeier *et al.*, 2011). Some conclude that international environmental agreements are effective (Levy, 1993; Victor *et al.*, 1998; Munton *et al.*, 1999; Wettestad, 2002a; Bratberg *et al.*, 2005). Others conclude that IEAs do not improve environmental performance beyond what could be achieved without them (Finus and Tjøtta, 2003; Ringquist and Kostadinova, 2005; Aakvik and Tjøtta, 2011). In fact, sometimes even studies on the very same institution have produced mixed or even contradictory evidence. The question if international environmental agreements are efficient means to protect the environment, and more specifically to reduce air pollution, appears necessary in particular against this background.

This study contributes to the literature on the effectiveness of IEAs (policies) in three ways. First, it takes the formidable hindrances to inference in observational studies seriously and aims at providing more credible estimates of the effects of IEA membership. Also, and contrary to previous studies, it lays out a detailed identification strategy and sets forth empirical tools for its implementation. It thus allows to model the non-institutional counterfactual, that is, the behavior of a member state in absence of institution membership. Second, it produces evidence on the effects of an air pollution agreement that has not yet been studied. It focuses on the Convention on Long Range Transboundary Air Pollution (CLRTAP), and more specifically on its latest institutional agreement, the Gothenburg Multi-Effect Protocol from 1999 which is one of the most prominent and long-standing examples of international environmental cooperation (UNECE, 1999).[2] Crafted in 1979, it currently consists of eight sequential protocols, each of them constituting a legally binding international agreement.[3] The LRTAP agreement is among the most comprehensive IEAs and is widely regarded as a particularly serious and sophisticated

[2]See the CLRTAP website of the United Nations Economic Commission for Europe (UNECE) for more information: `http://www.unece.org/env/lrtap/status/lrtaps.htm`.

[3]Following the framework convention concluded in Geneva in 1979, the LRTAP protocols address the following pollutants and emissions: Helsinki, 1985, SO_2 ; Sofia, 1988, NO_x ; Geneva, 1991, VOC (Volatile Organic Compounds); Oslo, 1994, SO_2 ; Aarhus, 1998, POPs (Persistent Organic Pollutants) and Heavy Metals; and Gothenburg, 1999, SO_2 , NO_x, VOC, NH_3 and three environmental effects.

attempt to solve an environmental problem at the international level and has itself attracted widespread attention in the scientific community (Murdoch *et al.*, 1997; Wettestad, 2002b; Finus and Tjøtta, 2003; Bratberg *et al.*, 2005; Ringquist and Kostadinova, 2005; Aakvik and Tjøtta, 2011). The Gothenburg 1999 protocol (UNECE, 1999) is voluntary and uses country specific emission targets which pundits believe to effectively influence states' pollution behavior towards reduction. This study is the first attempt to measure effectiveness of the Gothenburg protocol. Third, international institutions, whether they regulate environmental, labor or economic policy, often feature voluntariness and specificity of protocol targets. The Gothenburg 1999 protocol hence constitutes an ideal case to introduce and establish a tool-kit that promises to be more generally applicable to the empirical study of international institutions.

Measuring the Effectiveness of International Environmental Agreements: The Missing Counterfactual

This review of the literature focuses on the LRTAP agreement and the most relevant contributions to the study of environmental agreements more generally. Although there is a large body of qualitative studies on effectiveness, this study constrains itself to quantitative work, that is, studies that rely on numerical measurement and statistical methods. First, because this study itself constitutes a quantitative contribution, and second, the literature is too vast to review it completely here. Researchers employ two main strategies to empirically measure the effects of IEAs on state behavior. They either derive a numerical benchmark from game theoretic models (Helm and Sprinz, 2000; Hovi *et al.*, 2003) and/or in terms of theoretical conceptions of effectiveness scores (Helm and Sprinz, 2000; Miles *et al.*, 2002; Breitmeier *et al.*, 2006, 2011) and then estimate success versus these benchmarks, or they aim at estimating the causal effect of agreement membership (treatment) using multiple regression to account for problems due to the non-experimental nature of the data.[4] Both broader approaches are due to the inherent difficulty to justify what a state would have done if it did not join the institution. As this behavior involves the unobserved quantity of a potential outcome in absence of institutional

[4]To be clear, following the above definition, these studies are semi-quantitative in nature as they recur to some degree to qualitative assessments and transform them into numerical values before they apply statistical methods. Game theorists generally derive non-cooperative Nash-equilibria and/or social optima from their models as benchmarks. Also note that Miles *et al.* (2002), Breitmeier *et al.* (2006), and Breitmeier *et al.* (2011) evaluate variance in effectiveness across different international environmental agreements. This study restricts itself to measuring the effectiveness of a single agreement. This allows for identifying its causal effect a more transparent way.

membership, quantitative approaches evolve around strategies to model this *counterfactual*.

Deriving the Counterfactual from Theory

Helm and Sprinz (2000) construct a measure for effectiveness based on the ratio of actual institutional performance over optimal institutional performance. To get the nominator and denominator of this ratio, they subtract a 'non-regime' counterfactual from both the observed actual value of the institutions target variable and from the collectively optimal value of this variable respectively. The emerging *effectiveness score* constitutes a percentage deviation from theoretically optimal performance. In their study, they compute the 'non-regime' counterfactual from emission reductions predicted by a non-cooperative Nash-equilibrium using cost-benefit functions.[5] Analyzing the Helsinki 1985 protocol on SO_2 and the Sofia 1988 on NO_x they find that for both protocols the effectiveness scores are greater than and significantly different from zero.

Although this approach is theoretically elegant, it has at least three drawbacks. First, it is unlikely that the estimation strategy identifies the causal effect of the agreement, especially over time, since the agreement will potentially induce changes in states' cost-benefit functions. For example, according to regime theories, international agreements provide information at low costs, thereby altering the cost-benefit calculations of policymakers, which in turn leads to changes in state behavior (Keohane, 1984). Such effects would of course change the Nash predictions and thereby the estimated causal effect. Second, the values for the Nash equilibrium predictions stem from a set of predetermined factors. Whether these factors are valid empirically remains open to debate. Therefore, the appropriateness of the cost-benefit functions rests on a weak empirical foundation. Third, especially with complex cost-benefit functions, many Nash-equilibria can exist (Young, 2003). It remains unclear which of these multiple equilibria would constitute the correct counterfactual.

Model the Counterfactual using Multiple Regressions

A second group of studies employs multiple regression analysis to estimate causal effects of international environmental institutions. Mitchell and Deane (2009) use regressions to estimate

[5]The authors base their approach on Maler (1989).

the effect of three different LRTAP protocols, i.e., Helsinki 1985 (SO_2), Sofia 1988 (NO_x), and Oslo 1994 (the second SO_2 protocol), on state behavior. By including a set of covariates, including a lagged dependent variable, they intend to avoid omitted variable bias in the estimation. Mitchell and Deane (2009) find that none of the three protocols had a significant effect on the pollution behavior of states. They do report, however, that the first Helsinki 1985 protocol had an effect on state behavior for those participating states that are defined as *leader* states relative to non-participants.

Although the distinction between leader and laggard states is theoretically valuable and their findings here provide important empirical insights, the Mitchell and Deane (2009) model does not account for a self-selection into the treatment, i.e., the likely circumstance that states do not randomly become members of international environmental institutions. Put differently, institutional membership is not a random set. More specifically, self-selection occurs when a country bases its decision whether to join or to stay out of an international institution on a set of underlying (un-)observed factors. In the case of transboundary air pollution, for example, states that are highly affected by other states' emissions are likely to have a higher incentive to join an agreement since this might allow them to curb the impact of the respective pollutants. On the other hand, emitter states that do not suffer from significant health or ecological risks but only face huge abatement costs by having to implement new technical standards for polluting industries, are less eager to join an environmental agreement. Adding to this complexity of the data generating process, states' decisions are sometimes also based on less straightforward and hardly measurable factors. Perhaps, cultural factors overrule cost-benefit analyses and a government decides to join an institution even if it will face considerable costs. Some of these factors are directly observable, like the cost of technical standards, others are hard to quantify, e.g., ecological risks. As a result, treating institutional membership as a random set can lead us to either under- or overestimate the effectiveness of international environmental agreements. Because of this self-selective character, applying standard regression estimators can give rise to misleading and biased findings. Adding covariates to a regression can control for self-selection if the selection process only depends on the covariates specified in the particular model - this is unlikely in the most cases, though. In fact, it often deems inherently difficult to conclude if the chosen model does include the correct covariates and specifies the appropriate functional form. Hence, applying multiple regressions directly to the data casts doubt on the ability to construct a convincing counterfactual.

Simply using more controls cannot help to solve the problem either, since this means accept-

ing a higher risk that covariates are distributed in a way that can lead to a biased estimate. This follows from the higher probability that there are no overlaps between control units and treated units conditional on the covariates. This, in turn, means that inferences are made for value combinations of independent variables that do actually not exist in the data, but instead are based on extrapolations of existing values.[6] The cure might be worse than the disease then, since these extrapolations are usually based on assumptions about functional relationships between values of that variable in question. And precisely this assumption lacks justification again in most cases of real an applicable research. To illustrate this crucial point via the study's topic of interest: For the self-selection into institutional membership, we have to assume that covariates are unevenly distributed as countries decide their access on factors that differ between countries. Mitchell and Deane (2009) are aware of this selection problem, which they call *membership endogeneity*. They propose to include time-varying variables like *environmental vulnerability* and *abatement costs* to account for possible drivers of self-selection into the agreement. While this approach merits appreciation, still and as explained above, a relatively arbitrary proposition of variables does not necessarily account for selection bias.

Other studies take non-random assignment to institutions into perspective and offer econometric solutions to this problem. These studies account for systematic differences between treatment and control group, that is, between states that ratified the agreement and states that stayed outside the agreement. Ringquist and Kostadinova (2005) study the 1985 Helsinki protocol on SO_2 using a set of estimators that correct for non-random assignment into treatment. Additionally, they test the their models using pre-agreement time periods and a pre-treatment trend variable for the treated states, that is, agreement members, drawing on Heckman (1989). Their results suggest that joining the Helsinki 1985 protocol had no effect on a states' emissions. Ringquist and Kostadinova (2005) advance a random trend estimator on the basis of assuming that the systematic differences between treatment and control group with respect to the dependent variable are a function of state specific trends in that variable.

One key identifying assumption for the use of random trend models is that the state specific trend is independent from the regressors (Angrist and Pischke, 2008). Ringquist and Kostadinova (2005) unfortunately abstain from discussing the plausibility of this assumption. Additionally, although they carefully argue for a random trend estimator, they decide not to conduct Hausman tests to evaluate the choice of this estimator over a fixed effects estimator. Thus, their choice might lead to a efficiency loss. Furthermore, residuals for individual states

[6] Econometricians usually call this the *curse of dimensionality*.

are correlated over time periods if we apply a random trend estimator (Angrist and Pischke, 2008). Therefore, standard errors are usually estimated using generalized least squares (GLS). GLS in turn requires much stronger assumptions than ordinary least squares (OLS).[7] Ringquist and Kostadinova (2005) do not discuss these assumptions either. Hence, it remains at least unclear if their choice of estimator constitutes the most appropriate and efficient strategy to estimate the effect of the Helsinki 1985 protocol.

Model the Counterfactual using Causal Inference Approaches

The last group of studies draws on causal inference literature to take non-randomness of institutional membership into account. These studies aim at approximating randomization by study design. The identification of the causal effect to measure then gives rise to the appropriate empirical strategy.[8] In terms of identification strategy, Bratberg *et al.* (2005) and Aakvik and Tjøtta (2011) offer particularly sophisticated studies of the effectiveness of international environmental agreements. Both studies employ a Difference in Difference (DID) estimator. Bratberg *et al.* (2005) study the Sofia 1988 protocol on SO_2. They are the first to find an effect of the agreement on state behavior. Their study presents evidence of a 2.1% greater average reduction of SO_2 for treated states compared to their counterfactual response in absence of treatment. Moreover, they outline a comprehensive identification strategy.

While the study derives the key identifying assumption for the DID estimator, that is, parallel trends in potential outcomes, Bratberg *et al.* (2005) miss the opportunity to discuss its plausibility. Furthermore, but certainly less important, Bratberg *et al.* (2005) opt for first differencing the regression equation. Algebraically, differencing is the same as deviations from means, if and only if our data set expands over two periods. With more than two periods, deviations from means produces more efficient estimates (Angrist and Pischke, 2008). More importantly, by differencing, Bratberg *et al.* (2005) intend to control for time-constant characteristics that could drive selection into treatment. This can be problematic when selection into treatment depends on time-varying unobserved covariates. In the case of LRTAP technological innovation could be such a confounder.

In the most recent attempt to study effectiveness of an IEA, Aakvik and Tjøtta (2011) employ DID regressions to analyze the Helsinki 1985 and the Oslo 1994 protocols on SO_2. They find no significant effect of either the Helsinki nor the Oslo agreement in reducing SO_2 emissions.

[7] The covariance matrix of the errors has to be known.

[8] This study outlines the procedure in detail in the next chapter.

Their empirical modelling strategy addresses many of the above discussed problems. Although they prefer first differencing over deviations from means, as (Bratberg *et al.*, 2005), they introduce state specific time period dummies. These dummies capture time-varying effects of confounders in the sample. From a causal inference perspective this study forms the most convincing example of estimating the effect an international environmental agreement on states' emissions. Still, and somewhat surprisingly given the thoroughness of their approach, Aakvik and Tjøtta (2011) neither present an identification strategy that extends beyond naming the crucial parallel trends assumption, nor discuss the plausibility of this assumption.[9]

In the remainder of this paper, I outline a general identification strategy for estimating the causal effect of international environmental agreements, and in particular of the Gothenburg 1999 protocol on NO_x and SO_2 emissions on states' behavior, followed by discussion and empirical testing of the identifying (parallel trends) assumption. Then, I implement the estimation strategy in an empirical model that addresses the above mentioned challenges to identification.

Identification of the Causal Effect of Agreement Membership on State Behavior

Scholars that study the effects international environmental agreements confront the fundamental problem of causal inference: the impossibility of observing the counterfactual, i.e., the outcome for the same unit in the absence of the treatment. The ideal way to overcome this problem when trying to estimate the causal effect of an environmental institution would be to conduct an experiment in which agreement membership was randomly assigned to countries. Given random assignment, we could simply compare member states with non-member states. The difference between the average environmental performance measure of the treated (the agreement member countries) and the average of the respective indicator for the control group would constitute the causal effect of the agreement or protocol membership, because both groups are comparable with respect to (un-)observed confounders. Unfortunately, protocol membership is not randomly assigned to countries.[10] If confronted with non-random assignment, causal inference methods serve to overcome the obstacles to estimating causal effects (see e.g. Rubin, 1974, 1977; Holland, 1986; Angrist and Pischke, 2008).

[9] The next chapter elaborates on the importance of discussing identifying assumptions.

[10] Likewise, comparing pre- and post-treatment outcomes for the treated units most probably incorporate biases due to temporal trends in the outcome variable or to the effect of changes of other factors between periods (Abadie, 2004).

The identification strategy first defines its quantity of interest, the average treatment effect on the treated (δ_{ATET}), that is, the effect that protocol membership has on average on members of the protocol.[11] In terms of potential outcomes, we want to estimate:

$$\delta_{ATET} = E[Y_{1i}(1) - Y_{0i}(1)|D = 1], \tag{1}$$

where E[·] denotes expected value, $Y_{1i}(t)$ the potential outcome in periods for country i when treated, and $Y_{0i}(t)$ the potential outcome in periods for country i would have attained without treatment in time period t, that is, the counterfactual. This equation involves a big unobserved quantity. Formally, we do not observe the potential outcome:

$$E[Y_{0i}(1)|D = 1],$$

which is the outcome for untreated countries if they had been treated, i.e., the counterfactual. Following the notation of the Rubin Causal Model we assume that potential outcomes and treatment are independent, that is, assignment to treatment is independent of underlying unit characteristics, it is random:

$$Y_{1i}(t), Y_{0i}(t) \perp\!\!\!\perp D_{i,t}, \tag{2}$$

with $Y_{1i}(t)$ denoting the potential outcome at time t for country i when treated and $Y_{0i}(t)$ the potential outcome country i attains without treatment in time period t.

It is straightforward to see that random assignment implies that the expected value of potential outcomes for the treated under the treatment equals the expected value for the control if they had been treated, and likewise that the expected value for the treated if they had not been under treatment equals the expected value for the control that have not been treated:

$$E[Y_{1i}(t)|D = 0] = E[Y_{1i}(t)|D = 1] \quad \wedge \quad E[Y_{0i}(t)|D = 0] = E[Y_{0i}(t)|D = 1], \tag{3}$$

Given random assignment, we could simply look at the difference in means between the treated countries and the controls as this would measure the causal effect of the treatment. Now, as with international environmental agreements we do not have random assignment of countries to the treatment, i.e., to protocol membership, equation 3 does not hold. Put differently, contrary

[11] However, it is not clear if this is really the most useful approach. This is because recent research claims that IEAs do influence members as well as non members (Mitchell and Deane, 2009). So, there could also be an effect of the treatment on the control countries. Thus, this might be evidence of bias in our estimates. If we let aside this concern the quantity of interest is the δ_{ATET}.

to standard regression assumptions, treated and control states will likely differ on unobservable characteristics that are associated with potential outcomes even after controlling for differences in observed characteristics. Hence, under non-random assignment, looking at differences in means incorporates bias of the following form:

$$\begin{aligned} E[Y_i(t)|D=1] - E[Y_i(t)|D=0] &= E[Y_{1i}(t)|D=1] - E[Y_{0i}(t)|D=0] \\ &= \underbrace{E[Y_{1i}(1) - Y_{0i}(1)|D=1]}_{\delta_{ATET}} + \underbrace{(E[Y_{0i}(1)|D=1] - E[Y_{0i}(0)|D=0])}_{SelectionBias} \end{aligned}$$

One way to overcome this problem is the difference in difference estimator. The DID estimator rests on the idea that an untreated control group can be used to 'control' or in some sense 'remove' temporal variation in the outcome that is not due to the treatment. Applicability of the DID estimator hinges on the key identifying assumption of *parallel trends*:

$$E[Y_{1i}(1) - Y_{1i}(0)|D=1] = E[Y_{0i}(1) - Y_{0i}(0)|D=0], \tag{4}$$

which means that the average potential outcomes for the treated and the untreated units follow parallel trends over time. This allows the use of the observable difference in outcomes for the controls as the counterfactual for the treated.

Especially in the case of LRTAP protocols this assumption is very strong as treaty ratification is voluntary and we expect the treatment and the control group to show a different behavior in reduction paths over time. Hence, researchers need to carefully decide whether this assumption is plausible. Obviously, the parallel trends assumption does not allow empirical testing because it involves unobserved quantities, namely the time trend of the treated units in absence of the treatment.

However, their exists an approximative test to evaluate the plausibility of the parallel trends assumption (Angrist and Pischke, 2008, 238) . We can add state-specific time trends to the list of controls in the preferred empirical model.[12] This allows treatment and control states to follow different time trends. If the workhorse model produces unchanged coefficients on the interesting variables for both the specifications with and without state-specific time trends, this adds to the credibility of the parallel trends assumption. Given parallel trends, equation (5)

[12]To make this explicit, we can multiply a state specific trend coefficient with the time trend variable (Angrist and Pischke, 2008).

holds and we attain identification of average treatment effect (δ_{ATE}):

$$\begin{aligned}\delta_{ATE} &= E[Y_{1i}(1) - Y_{0i}(1)|D = 1] \\ &= \{E[Y(1)|D = 1] - E[Y(1)|D = 0]\} - \{E[Y(0)|D = 1] - E[Y(0)|D = 0]\}, \end{aligned} \tag{5}$$

where δ_{ATE} is the average treatment effect, $Y(1)$ is the average outcome under treatment, and $Y(0)$ the average outcome before treatment.[13] From (5) and (1), it is easy to see that in this case

$$\delta_{ATE} = \delta_{ATET},$$

which is the expression we specified as quantity of interest in the beginning. Now, we can construct an estimator to measure δ_{ATET} (Abadie, 2004):

$$\begin{aligned}&\left\{\frac{1}{n_1}\sum_{D_i=1} Y_i(1) - \frac{1}{n_0}\sum_{D_i=0} Y_i(1)\right\} - \left\{\frac{1}{n_1}\sum_{D_i=1} Y_i(0) - \frac{1}{n_0}\sum_{D_i=0} Y_i(0)\right\} \\ &= \left\{\frac{1}{n_1}\sum_{D_i=1} \{Y_i(1) - Y_i(0)\} - \frac{1}{n_0}\sum_{D_i=0} \{Y_i(1) - Y_i(0)\}\right\},\end{aligned}$$

where n_0 and n_1 are counters for control and treatment group states respectively. In the case of the LRTAP, observed covariates, e.g., ecological vulnerability or population size, might influence time dynamics for member and non-member states.[14] Using a linear regression framework with covariates allows us to control for these dynamics[15]

$$\Delta Y = \alpha + X'\beta + \delta D + u, \tag{6}$$

where α is a constant, X' a vector of covariates, D a dummy for membership in the institution, and u an error term. Furthermore, in the case of international environmental agreements, selection into an institution might be driven by unobserved time-constant characteristics of states, e.g. political culture or, even more prominent in the case of sequentiality of protocols like in the LRTAP, previous membership in other protocols. Thus, the study employs the DID estimator in a fixed effects regression framework

$$Y_i(t) = \eta_i + \gamma(t) + X_i'\beta(t) + \delta D_i(t) + \varepsilon_i(t), \tag{7}$$

[13] Note that we look at actual outcomes now.

[14] Apart from controlling for compositional effects, adding covariates can enhance statistical precision.

[15] At least as long they can be linearly explained. The precision of our DID estimator is directly given by the standard error of δ

where η_i is a time-constant state fixed effect possibly correlated with D_i, $\gamma(t)$ are period effects, and $\varepsilon_i(t)$ is the error term. Furthermore, the study employs time period fixed effects to take into account effects that may influence all cases in a given time period to the same amount. This might be a technological shock at the macro level that is available to all countries, or a macro business cycle shock that influences all states cost-benefit analyses similarly.[16]

Data and Variables

Dependent Variable and Treatment Indicator

Environmental effectiveness variables can be categorized into output, outcome and impact variables (Underdal, 2008). Output refers to the formal implementation of institutional targets, or compliance. Outcome denotes the change in human behavior associated with the international environmental agreements, for example a 10 percent reduction in a country‘s annual SO_2 emissions. Impact concerns the actual change of the biophysical environment, for example an improvement in air quality measured in terms of particulate matter as a percentage of air volume. The study's empirical methodology allows for the analysis of virtually every environmental effectiveness variable and is thus effortlessly applicable to evaluate different international environmental institutions.

Ultimately, this study intends to analyze whether air pollution agreements really improve air quality. Therefore, it focuses on outcome or impact indicators which measure environmental quality (Mitchell, 2008). The accuracy of these measures depends on measurement techniques and, in the case of air pollution, on how external influences such as wind or water transportation patterns are taken into account. In the case of air pollution agreements, these indicators are well developed and relatively trustworthy (EMEP, 2008b,a).

Analyzing outcome measures, such as NO_x and SO_2 is particularly useful because of two reasons. First, the Gothenburg protocol requires ratifying states to reach state-specific ceilings by 2010 formulated as percentage emission reductions relative to the 1980 annual emissions levels.[17] Second, the causal chain between the LRTAP agreement and any given impact indi-

[16] However, fixed effects regression works poorly if selection into treatment depends on time varying covariates. In the case of LRTAP, individual state level technological innovation could be such a confounder. Still, the DID approach is the most rigorous way to estimate institution effectiveness in the case of LRTAP.

[17] Although all protocols are voluntary by nature, they differ on the calculation of emission targets. First, the framework convention from 1979 consists of several loosely formulated emission targets. Second, the 1985, 1988 and 1991 protocols formulated uniform reduction targets for all ratifying parties. Finally, the Oslo protocol from 1994 introduced the *critical loads approach* which applies individual and varying reduction targets to

cator is likely to be too long to capture the actual effect of the institution. Mitchell (2008, 84) convincingly argues that the longer the causal chain from the environmental indicator to the respective agreement, the more alternative factors could explain changes in the given environmental indicator. Hence, this study uses NO_x and SO_2 emissions per country as it is the most closest measure of the quantity that should ultimately change.

Time series of emission levels are notoriously non-stationary, that is, residuals exhibit strong serial auto-correlation. First differencing can solve this problem. Unit root tests first reject the null hypothesis of stationarity for SO_2 emissions.[18] After differencing the test rejects the hypothesis of all panels having unit roots both for SO_2. Thus, the dependent variable is first differenced logarithms of annual emission levels for SO_2 per country, i.e., growth rates. In the case of NO_x, unit root tests confirm stationarity. The study thus performs the analysis with NO_x emission levels. The data set includes 43 European and Eurasian countries from 1995 to 2008.[19] The European Monitoring and Evaluation Program (EMEP) provides emission data on its homepage.[20] The data is complete over the whole time-series for all countries in the data set. The treatment indicator, *Ratification*, is a binary variable that takes the value 1 if country i is a member of the Gothenburg protocol in year t (ratified), and 0 else. Hence, a country that ratifies the protocol between 1995 and 2008 belongs to the treatment group, respectively to the control group if it does not ratify the protocol.

Figure (1) shows yearly average growth rates of NO_x and SO_2 emissions for control and treatment group. Growth rates illustrate group specific behavior better than absolute values. Obviously, member states demonstrate larger reduction rates than non-members. This can be interpreted as a sign of selection into treatment based on group characteristics. Both control and treatment group countries experience roughly similar trends in pre-treatment years.[21] However, mean growth rates of emissions during the period in which the protocol was ratified prevent visual interpretation as both treatment and control countries show no clear pattern on average. This holds for both pollutants. This purely descriptive illustration does not support the view that the protocol has a positive effect on states' pollution reducing behavior.

Non-parametric DIDs for both substances serve as a next point in descriptive analysis of aver-

account for country differences. The Gothenburg 1999 protocol continues this approach.

[18] All test results available from the author.

[19] See Table (10) in the appendix for a detailed list of countries in the sample and their ratification behavior.

[20] www.emep.int. EMEP is scientifically based and policy driven, and thus widely believed to be politically independent.

[21] Note that visual inspection functions as a first check point to evaluate the parallel trends assumption. The exception of year 2000 in NO_x emission growth rates might appear extreme, but note that mathematically the difference constitutes roughly 0.15%.

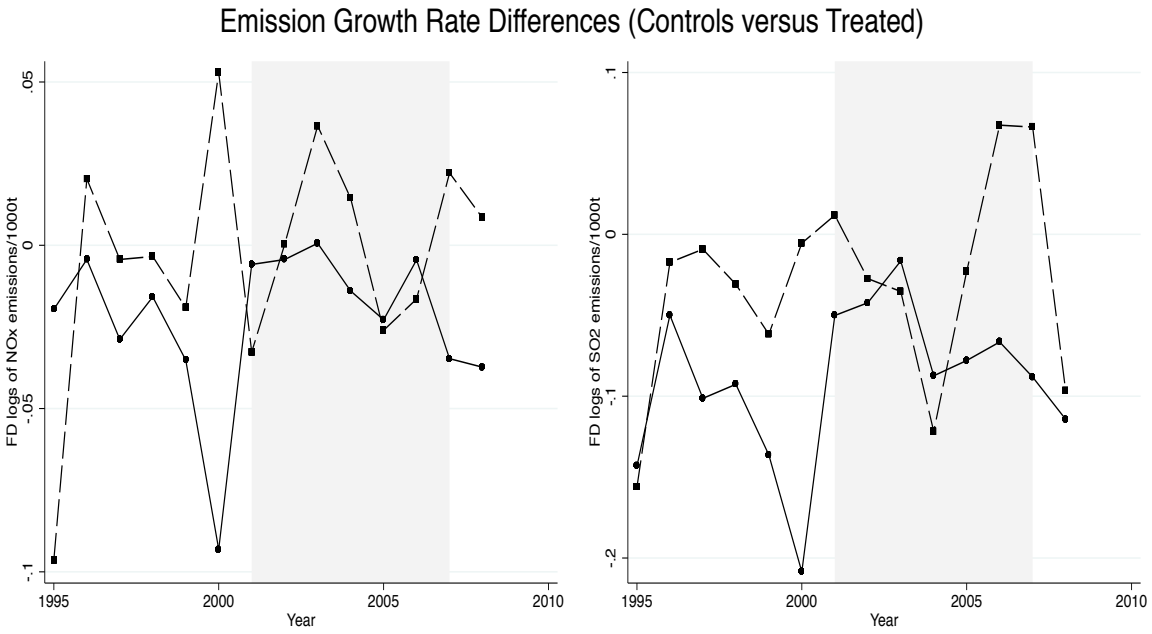

Figure 1: Average annual growth rates for logs of NOx and SO2 emissions. The solid line shows states that become protocol members at t. The dashed line shows permanent non-member states. Shaded area marks range of years in which states ratified the protocol.

age time series evolution. Figure (2) and (3) show relative annual differences in growth rates between control and treatment group countries for three distinctive time periods in which the respective countries ratified the protocol. Analytically, these figures merit attention because they depict the differences between treatment and control group using the same approach as in the empirical part but without accounting for confounding factors, e.g., mediating factors or unit specific time constant effects. Hence, the figures show a first, albeit rough, description of the possible effect that is likely to be measured afterwards. Consequently, we expect negative differences in growth rates in case the protocol might affect states' behavior on average. For NO_x emissions, treated countries demonstrate this behavior relative to controls in the period 2006 and 2006, and also, after a positive difference in 2001, for the years 2004 and 2005. Relative differences in growth rates for the years 2004, 2005 counterintuitively are positive. Behavior in relative SO_2 emission growth rates is less encouraging. For 2001 to 2003, country groups' relative growth rates are zero or positive, for 2004 to 2005 negative only in the second period and for 2006 to 2007 negative and zero, respectively. Descriptively, growth rate differences point to non-effectiveness of the protocol then.

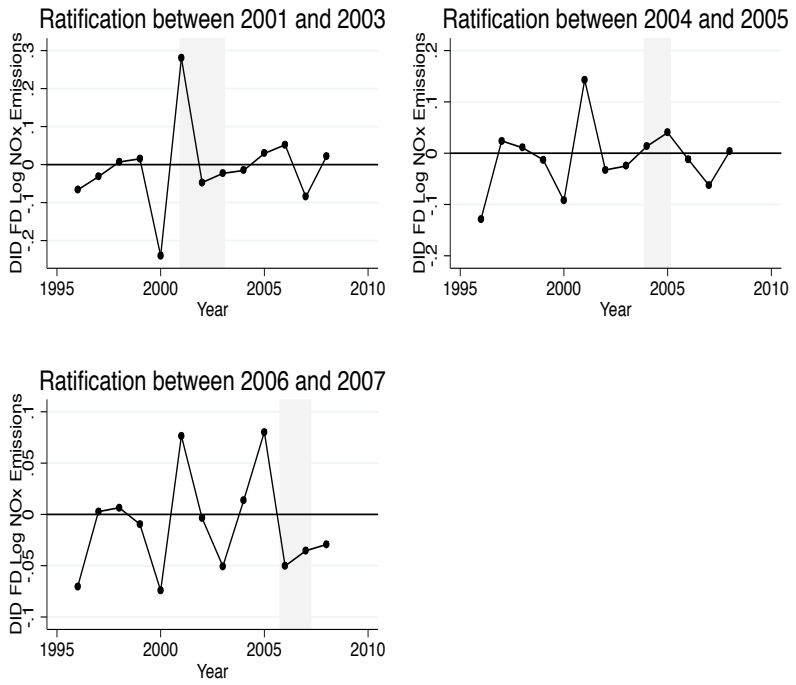

Figure 2: Non-parametric yearly DID for logs of NO_x emissions for country groups. Negative values mean higher reduction rates in treatment versus control group countries on average. Positive values mean higher reduction rates in control group versus treatment countries on average.

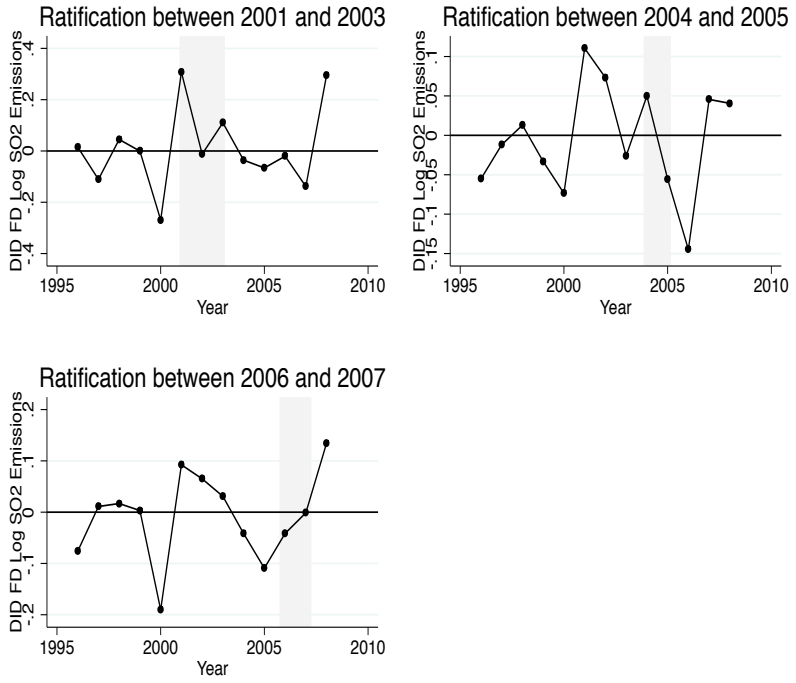

Figure 3: Non-parametric yearly DID for logs of SO_2 emissions for country groups. Negative values mean higher reduction rates in treatment versus control group countries on average. Positive values mean higher reduction rates in control group versus treatment countries on average.

Covariates

A first covariates is receiver-emitter matrices where a country's total emissions are disaggregated to self-depositions, imports and exports of a pollutant (EMEP).[22] The matrices are available from 1985 onwards.[23] The study uses NO_x self-depositions and imports, NO_x *Selfdepositions* and NO_x *Imports*, as predictors for a country's pollution emitting behavior. While high self-depositions might drive countries to abstain from ratifying an agreement, high imports of the same pollutant might act as an incentive to engage in international environmental institutions. Thus, both factors might further selection effects. Another predictor of NO_x emissions is environmental vulnerability that is proxied by the share of a country's territory covered with forest, *Forest*. Data for forest coverage is available in five year periods. The study linearly interpolates missing data for the forest variable. Interpolation seems straightforward as forest cover is neither very volatile nor usually experiencing random shocks in between the measurement points.

Finally, the gross domestic product per capita, *GDP p.c.*, and the size of the population, *Population*, also serve as predictors for a country's behavior in terms of NO_x emissions. It is noteworthy that GDP is a per capita measure to capture a country's productivity which is positively correlated with higher emissions, and the population measure is an absolute number that exhibits the same positive relationship. Data for these variables is complete over the whole time-series for all countries and can be obtained from the World Bank's World Development Indicators from 2010.[24] The study discusses the treatment of missing data in the appendix.

[22] A tabular overview of variables, their descriptions, and data sources can be found in the appendix under (6).
[23] Matrices from 1997 onwards are available on the EMEP homepage, EMEP, earlier years from Sandnes (1993).
[24] Summary statistics for all variables can be found in the appendix on table (7).

Results and Sensitivity Analysis

Main Results for NOx Emissions

Table (1) presents the results for the DID with fixed effects with clustered standard errors on logs of NO_x emission levels.[25] As explained in the methodological part, the identifying parallel trends assumption undergoes a plausibility check. In the appendix, table (8) shows encouraging evidence for NO_x emissions. Interpretation of the treatment indicator effect on the dependent variable is straightforward as the study uses logs of NO_x emissions. The causal effect of protocol ratification on state behavior thus reads as a percentage change in emission levels. Remember that the identified effect is the average treatment effect on the treated. In the baseline model (1), a fixed effects regression on the treatment indicator and a time trend variable, ratification causes NO_x emissions to plunge with a annual average rate of 14%. However, if the model (2) incorporates additionally covariates, the effect vanishes. These covariates moderate observable heterogeneity between control and treatment group then. The preferred model (3) additionally includes year fixed effects that capture time variant heterogeneity that has the same effect on all states.[26]

The results suggest that protocol membership has no effect. Although the coefficient on ratification has the expected negative sign, it is not significant at any conventional level. Significant coefficients on self-deposited NO_x, forest cover, and GDP per capita show positive signs. It is important to bear in mind that these control variables have no causal interpretation. Because of the parallel trends assumption they can only be interpreted as mediating factors for the treatment effect.[27] Still, we can take a closer look at these mediating factors. As we would expect, high self-depositions and high emissions are correlated. If dependent and independent variables are log transformed, coefficients read as constant elasticities. Thus, a 1% increase in NO_x self-depositions is associated with a 0.11% increase in NO_x emissions. The same holds for GDP per capita. A 1% increase in GDP per capita is associated with a 0.26% increase in NO_x emissions. This comes not unexpectedly as countries who produce more often pollute more. Rather counterintuitively the extent of forest cover is also positively associated with

[25] A random effects model would have been another possibility to estimate the quantity of interest (Wooldridge, 2005). A Hausman test rejects the random effects model in favor of the fixed effects model. This holds for both SO_2 and NO_x. Results available from the author.

[26] Plotted time series of average log emissions suggest not to add quadratic trends to the model. Adding quadratic terms to a functionally linear model could instead introduce misspecification bias.

[27] This in turn is convenient for the researcher because covariates do not need to be theoretically well justified or what is even more difficult and oftentimes quite arbitrary, be interpreted theoretically in post-estimation analysis.

NO_x emissions.[28] The size of the population and imported NO_x emissions are not statistically different from zero. A 1% larger forest cover is correlated with an average 1.47% more NO_x emissions. Having said that, ratification of the Gothenburg 1999 protocol has no independent effect on states' emission behaviors with regard to NO_x emissions.

Table 1: Logs of NOx Emissions in 1000t/year

	(1)	(2)	(3)
Ratification	-0.14**	-0.04	-0.07
	(0.06)	(0.05)	(0.07)
NOx Selfdepositions		0.12**	0.11**
		(0.05)	(0.05)
NOx Imported		0.04	0.07
		(0.10)	(0.11)
Forest		1.46**	1.47**
		(0.55)	(0.56)
GDP p.c.		0.29***	0.26**
		(0.11)	(0.11)
Population		0.52	0.46
		(0.72)	(0.74)
Constant	13.32	52.87***	57.63***
	(13.81)	(12.71)	(13.00)
Time Trends	x	x	x
Year Fixed Effects			x
N	601	565	565

Note: Regression coefficients with robust standard errors next to coefficients (standard errors are clustered by country). Stars indicate significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Models 1-3 are fixed effects regressions where the dependent variable is the log of NOx emissions in 1000t per year.

[28] Ringquist and Kostadinova (2005) also find a positive, yet insignificant, correlation between forest cover and SO_2 emissions for the Helsinki 1985 protocol.

Main Results for SO2 Emissions

Table (2) presents the results for the DID with fixed effects with clustered standard errors on first differenced logs of SO_2 emission levels. Again, the study offers a plausibility check for the identifying parallel trends assumption. In the appendix, table (9) shows encouraging evidence for SO_2 emissions. Interpretation of the treatment indicator is somewhat more complicated than in the NO_x case. The causal effect of ratification on state behavior means here a percentage change on SO_2 growth rates. Suppose the coefficient on ratification is -0.02 and significant, then ratifying the protocol causes the average state to reduce emission growth rates about two percent. For SO_2, the baseline model (1) cannot identify an independent effect of agreement ratification on state behavior. Adding covariates in model (2) does not change these results. The preferred model (3) with country and year fixed effects again reaffirms the results from models (1) and (2). Hence, the treatment indicator is not significant at any conventional significance level. Ratification of the Gothenburg 1999 protocol does not cause the average member state to change its polluting behavior with regard to SO_2 emissions. Although covariates show the expected signs, with the exception of imported SO_2, they are all together not significant at conventional significance levels.[29]

[29]Exclusion of covariates did not change the results substantially.

Table 2: First Differenced Logs of SO2 Emissions in 1000t/year

	(1)	(2)	(3)
Ratification	0.01	0.04	0.04
	(0.03)	(0.03)	(0.03)
SO2 Selfdepositions		0.02	0.04
		(0.02)	(0.03)
SO2 Imported		-0.04	-0.08
		(0.05)	(0.06)
Forest		-0.16	-0.16
		(0.20)	(0.21)
GDP p.c.		0.09	0.11
		(0.08)	(0.08)
Population		0.09	0.13
		(0.47)	(0.47)
Constant	-5.76	4.59	9.70
	(5.53)	(7.48)	(12.61)
Time Trends	x	x	x
Year Fixed Effects			x
N	600	565	565

Note: Regression coefficients with robust standard errors next to coefficients (standard errors are clustered by country). Stars indicate significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Models 1-3 are fixed effects regressions where the dependent variable first differences of the log of SO2 emissions in 1000t per year.

Sensitivity Analysis

Although, so far, the results suggest that protocol ratification is not effective on a state's emission producing behavior, concluding that the Gothenburg 1999 protocol has no impact would be premature. The empirical analysis might be sensitive to model specification or to misconceptualization of the treatment indicator. The study offers sensitivity analyses that addresses these concerns.

Individual trends in unobserved country characteristics sometimes confound estimation results. The inclusion of parametric country specific time trends possibly absorbs such time variant heterogeneity.[30] Table (3) presents estimation results for the preferred models for both NO_x, model (1), and SO_2, model (2), emissions. For both outcome variables, the model specification has no impact on the results.[31] The treatment indicator stays insignificant at conventional significance levels.

The second concern might emanate from the conceptualization of the treatment indicator. Specifically, the timing of the treatment could confound the estimation results. State governments might anticipate ratification and take action before the actual date of ratifying the agreement. Contrary, even if the agreement is ratified, observing reduced emission levels might be delayed because of hindrances in the implementation process, such as political or technological issues. To incorporate such possible anticipation or delay effects, the study runs placebo tests where the treatment changed to two years earlier and two years after the actual date of ratification. Table (4) and table (5) present placebo tests for NO_x and SO_2 emissions respectively. Models (1), (2), and (3) are based on the treatment taking place two years in advance, models (4), (5), and (6) on the treatment two years after actual ratification of the protocol. As before, model (3) contains the preferred specification. Obviously, results stay virtually unchanged for all specifications. Sensitivity analyses thus conform non-effectiveness of the Gothenburg 1999 protocol on NO_x and SO_2 emissions.

[30] Ideally, the model specifies non-parametric country specific time trends. This sample, though, does not contain enough observations.

[31] Sensitivity analysis does not include the estimation of first differenced equations. If idiosyncratic errors are not serially correlated, fixed effects estimation is relatively more efficient (Wooldridge, 2001). As mentioned before, tests confirmed panel stationarity. I still ran the first difference models. Results are robust and can be obtained from the author.

Table 3: Country Specific Time Trends

	(1)	(2)
Ratification	-0.00	0.03
	(0.04)	(0.04)
NOx Selfdepositions	0.02	
	(0.03)	
NOx Imported	0.10	
	(0.08)	
SO2 Selfdepositions		0.01
		(0.03)
SO2 Imported		-0.05
		(0.08)
Forest	4.35**	-0.01
	(1.75)	(1.99)
GDP p.c.	0.11	0.03
	(0.16)	(0.15)
Population	-0.49	0.89
	(0.97)	(1.36)
Constant	75.56***	-5.67
	(23.09)	(20.14)
Country Spec. Time Trends	x	x
Year Fixed Effects	x	x
N	565	565

Note: Regression coefficients with robust standard errors next to coefficients (standard errors are clustered by country). Stars indicate significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Model 1 is a fixed effects regression with country specific time trends for both controls and treated and year fixed effects where the dependent variable is the log of NOx emissions in 1000t per year. Model 2 is the same specification where the dependent variable are first differences of the log of SO2 emissions in 1000t per year.

Table 4: Placebo Tests Pre- and Post-Treatment for NOx

	(1)	(2)	(3)	(4)	(5)	(6)
Ratification	-0.14**	-0.03	-0.04	-0.15***	-0.07	-0.09
	(0.06)	(0.05)	(0.06)	(0.05)	(0.04)	(0.05)
NOx Selfdepositions		0.12**	0.11**		0.12**	0.11*
		(0.05)	(0.05)		(0.05)	(0.06)
NOx Imported		0.05	0.07		0.04	0.07
		(0.10)	(0.11)		(0.10)	(0.11)
Forest		1.47**	1.48**		1.46**	1.49**
		(0.55)	(0.56)		(0.55)	(0.56)
GDP p.c.		0.30**	0.27**		0.30***	0.27**
		(0.11)	(0.11)		(0.10)	(0.11)
Population		0.53	0.48		0.53	0.47
		(0.73)	(0.74)		(0.69)	(0.70)
Constant	11.82	54.74***	60.29***	16.96	53.16***	57.57***
	(14.67)	(14.05)	(13.17)	(12.30)	(11.59)	(12.29)
Time Trends	x	x	x	x	x	x
Year Fixed Effects			x			x
N	601	565	565	601	565	565

Note: Regression coefficients with robust standard errors next to coefficients (standard errors are clustered by country). Stars indicate significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Models 1-3 are placebo regressions replicating original model specifications with the treatment shifted country-wise two years backwards. Models 4-6 shift the treatment country-wise two years forwards.

Table 5: Placebo Tests Pre- and Post-Treatment for SO2

	(1)	(2)	(3)	(4)	(5)	(6)
Ratification	-0.02	-0.00	-0.01	0.02	0.03	0.04
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
SO2 Selfdepositions		0.02	0.03		0.02	0.03
		(0.02)	(0.03)		(0.02)	(0.03)
SO2 Imported		-0.04	-0.08		-0.04	-0.07
		(0.05)	(0.06)		(0.05)	(0.06)
Forest		-0.19	-0.19		-0.18	-0.17
		(0.22)	(0.22)		(0.20)	(0.21)
GDP p.c.		0.08	0.08		0.08	0.10
		(0.08)	(0.08)		(0.08)	(0.08)
Population		0.06	0.09		0.07	0.12
		(0.46)	(0.46)		(0.46)	(0.47)
Constant	-8.96	0.30	3.69	-6.04	2.15	8.09
	(6.18)	(10.31)	(14.72)	(5.18)	(7.72)	(11.63)
Time Trends	x	x	x	x	x	x
Year Fixed Effects			x			x
N	601	565	565	601	565	565

Note: Regression coefficients with robust standard errors next to coefficients (standard errors are clustered by country). Stars indicate significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Models 1-3 are placebo regressions replicating original model specifications with the treatment shifted country-wise two years backwards. Models 4-6 shift the treatment country-wise two years forwards.

Conclusions

Environmental institutions are prominent examples of international cooperation. Over the past decades, states increasingly turned to international institutions to protect the environment. However, theoretical and empirical debates about effectiveness of international environmental institutions are controversial. Existing approaches to measure the effectiveness of international environmental institutions have encountered various obstacles to conceptualize a convincing counterfactual, that is, the behavior of institutional members if they stayed outside the institution. This study proposes a tool-kit to estimate the effectiveness of institutional membership on states' behavior. It takes non-random assignment to the treatment, that is, institutional membership, explicitly into account and suggests a Difference in Difference estimator to control for selection bias. In doing so, it discusses how to identify the causal effect of protocol ratification on the average member and sets up a transparent model specification. Additionally, it argues for justification of identifying assumptions, in the case of DID, the parallel trends assumption, and provides an approximation test. It thereby adds to the literature on institutional effectiveness in general and presents an approach to empirically study the effectiveness of international institutions.

Empirically, this study uses NO_x and SO_2 emission data for 43 European countries to estimate the causal effect of the Gothenburg 1999 protocol on states' emission reduction behavior. It finds that protocol membership had no significant effect on state behavior, neither in the case of NO_x nor SO_2 emissions. The study thus strengthens arguments that predict ineffectiveness of international environmental institutions. Apparently, the similarity of the protocol to other international environmental institutions supports this conclusions. The question therefore arises if international environmental institutions in general only codify preexisting tendencies in state behavior. If this is true, they only mark a point where international cooperation is cheap in the sense, that after ratification states follow their behavioral patterns just as before. Thus, international environmental institutions may read as signals for environmental issues that are associated with non-conflicting state interests.

Finally, some limitations of the study have to be pointed out. The scope of this paper limited its analysis to present tools to identify average effects on member states after agreement ratification and to provide evidence for a single international environmental institution that regulates two substances. Although the Gothenburg 1999 protocol is comparable in important ways to other environmental institutions, the results do not necessarily generalize to them.

Other institutions may effectively change state behavior. The study thus might serve as a point in case for the search for different approaches to regulate state behavior internationally, it represents no carte blanche to reject international institutional solutions, though. If we assume that the results of this study translate to other institutions, should we now conclude that international environmental agreements are a waist of resources? That would most certainly constitute a premature conclusion but, equipped with the presented tool-kit, further analyses of international environmental institutions could help to construct a more illustrative picture of institutional effectiveness in general.

Also, further analyses should amplify scope in two ways. First, by considering different points in the negotiation process as possibly having an impact on state behavior. Traditionally, studies of international environmental institutions focus on ratification as the theoretically important point in a negotiation process. Work by Stein (2005) and Simmons and Hopkins (2005) theorizes about possible anticipation effects of international institution membership. Hence, maybe not ratification but signature of an agreement causes effects on states. Follow up work by this author builds on anticipation arguments and investigates the possible effects of signature on state behavior. Second, effects might not exist on average, but for individual states or specific state groups the picture could look different. Mitchell and Deane (2009) clearly provide interesting theoretical considerations regarding that issue. Hopefully, this study encourages further research along these lines.

Appendix

Treating Missing Values

I deleted a set of post-sovjet countries from the data-set because of missing data, namely Kyrgyzstan, Turkmenistan, and Tajikistan. Liechtenstein was dropped for the same reason. The study does not use interpolated data on the dependent variable because the data shows no linear pattern. For Russia the data appeared to be inconsistent, i.e., data from the EMEP model and the receiver-emitter matrices differed so highly that they were not considered trustworthy. Serbia and Montenegro are treated as one because they only split in 2006. Canada and the USA are excluded because they are not geographically contiguous with the other countries and therefore face, from a geophysical viewpoint, not the same transboundary constraints as the other countries.

Table 6: List of Variables

Dependent variables	First differenced Log of NO_x emission in 1000 tonnes per country i in year t. Source: EMEP `http://www.ceip.at/emission-data-webdab/`
	First differenced Log of SO_2 emission in 1000 tonnes per country i in year t. Source: EMEP `http://www.ceip.at/emission-data-webdab/`
Ratification	Treatment dummy
NO_x **Selfdepositions**	Log of total self-depositions on a country from total of emitted NO_x by the same country by year in 1000 tonnes. Source: (EMEP, 2008a,b)
NO_x **Imports**	Log of total depositions of NO_x from other countries in a country by year in 1000 tonnes. Source: (EMEP, 2008a,b)
Forest	Log of forest cover of a country in percent of total territory, interpolated. Source: (World-Bank, 2010)
GDP p.c.	Log of gross domestic product per capita. Source: (World-Bank, 2010)
Population	Log of total number of population per country. Source: (World-Bank, 2010)

Table 7: Summary statistics

		Non-Member States			
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Log NOx	4.7	1.4	2.1	7.5	279
Differenced log SO2	0	0.2	-0.8	1.6	278
NOx Selfdepositions	3.3	2.2	-1.9	7	268
NOx Imported	5.5	1.5	0.7	8	268
SO2 Selfdepositions	4.9	2	0.7	8.8	268
SO2 Imported	6.4	1.5	1.6	9.3	268
Forest	2.7	1.5	-2	4.2	280
GDP p.c.	8.1	1.4	5.8	10.5	260
Population	15.6	1.5	12.5	18.1	266
		Member States			
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Log NOx	5.3	1.3	2.6	7.8	322
Differenced log SO2	-0.1	0.2	-1.2	0.9	322
NOx Selfdepositions	3.7	1.8	0	7.1	318
NOx Imported	5.9	1.2	2.3	8	319
SO2 Selfdepositions	4.7	1.8	0	8.2	318
SO2 Imported	6.4	1.2	2.7	8.9	319
Forest	3.4	0.5	2.3	4.3	322
GDP p.c.	9.4	1	7.2	10.9	321
Population	15.9	1.3	12.9	18.2	322

Note: Variables are logarithms. Differenced SO_2 emissions read as annual growth rates.

Table 8: Testing the Parallel Trends Assumption for NOx Emissions

	(1)	(2)	(3)
Ratification	-0.14**	-0.07	-0.02
	(0.06)	(0.07)	(0.03)
NOx Selfdepositions		0.11**	0.07*
		(0.05)	(0.04)
NOx Imported		0.07	0.07
		(0.11)	(0.06)
Forest		1.47**	0.91*
		(0.56)	(0.53)
GDP p.c.		0.26**	0.13
		(0.11)	(0.17)
Population		0.46	-0.78
		(0.74)	(0.69)
Constant	13.32	57.63***	59.17***
	(13.81)	(13.00)	(8.68)
State-Specific Time Trends			x
N	601	565	565

Note: Regression coefficients with robust standard errors next to coefficients (standard errors are clustered by country). Stars indicate significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Model 1 is a fixed effects regression where the dependent variable are logs of NOx emissions in 1000t per year. Model 2 is the fixed effects regression with covariates. Model 3 adds state-specific time trends to the control group.

Table 9: Testing the Parallel Trends Assumption for SO2 Emissions

	(1)	(2)	(3)
Ratification	0.01	0.04	0.05
	(0.03)	(0.03)	(0.05)
SO2 Selfdepositions		0.04	0.01
		(0.03)	(0.03)
SO2 Imported		-0.08	0.01
		(0.06)	(0.10)
Forest		-0.16	-0.23
		(0.21)	(0.81)
GDP p.c.		0.11	-0.10
		(0.08)	(0.26)
Population		0.13	0.02
		(0.47)	(1.21)
Constant	-5.76	9.70	-7.93
	(5.53)	(12.61)	(12.77)
State-Specific Time Trends			x
N	600	565	565

Note: Regression coefficients with robust standard errors next to coefficients (standard errors are clustered by country). Stars indicate significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Model 1 is a fixed effects regression where the dependent variable is the first difference of logs of SO2 emissions in 1000t per year. Model 2 is the fixed effects regression with covariates. Model 3 adds state-specific time trends to the control group.

Table 10: Classification of States and the Gothenburg 1999 Multi-Effect Protocol

Country	**Protocol Signature**	**Protocol Ratification**	**CLRTAP Ratification**	**Control/Treatment Group**
Albania	no	no	2005	Control
Armenia	1999	no	1997	Control
Austria	1999	no	1982	Control
Azerbaijan	no	no	2002	Control
Belarus	no	no	1980	Control
Belgium	2000	2007	1982	Treatment
Bosnia and Herzegovina	no	no	1992	Control
Bulgaria	1999	2005	1981	Treatment
Croatia	1999	2008	1992	Control
Cyprus	no	2007	1991	Treatment
Czech Republic	1999	2004	1993	Treatment
Denmark	1999	2002	1982	Treatment
Estonia	no	no	2000	Control
Finland	1999	2003	1981	Treatment
France	1999	2007	1981	Treatment
Georgia	no	no	1999	Control
Germany	1999	2004	1982	Treatment
Greece	2000	no	1983	Control
Hungary	1999	2006	1980	Treatment
Iceland	no	no	1983	Control
Ireland	1999	no	1982	Control
Italy	1999	no	1982	Control
Kazakhstan	no	no	2001	Control
Latvia	1999	2004	1994	Treatment
Lithuania	no	2004	1994	Treatment
Luxembourg	1999	2001	1982	Treatment
Macedonia	no	2010	1997	Control
Malta	no	no	1997	Control
Moldova	2000	no	1995	Control
Netherlands	1999	2004	1982	Treatment
Norway	1999	2002	1981	Treatment
Poland	2000	no	1985	Control
Portugal	1999	2005	1980	Treatment
Romania	1999	2003	1991	Treatment
Serbia and Montenegro	no	no	2001	Control
Slovakia	1999	2005	1993	Treatment
Slovenia	1999	2004	1994	Treatment
Spain	1999	2005	1982	Treatment
Sweden	1999	2002	1981	Treatment
Switzerland	1999	2005	1983	Treatment
Turkey	no	no	1983	Control
Ukraine	no	no	no	Control
United Kingdom	1999	2005	1982	Treatment

Source: UN ECE, LRTAP Convention, 08. November, 2010.

References

Aakvik, A. and Tjøtta, S. (2011). Do collective actions clear common air? the effect of international environmental protocols on sulphur emissions. *European Journal of Political Economy*, **27**(2), 343–351.

Abadie, A. (2004). Causal inference. *Encyclopedia of Social Measurement*, **00**, 1–8.

Andreaevc, M. O. and Crutzen, P. J. (1997). Atmospheric aerosols: Biogeochemical sources and role in atmospheric chemistry. *Science*, **276**(5315), 1052–1058.

Angrist, J. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

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

Bernauer, T. (1995). The effect of international environmental institutions: how we might learn more. *International Organization*, **49**(2), 351 – 377.

Böhmelt, T. and Pilster, U. (2010). International environmental regimes - legalisation, flexibility and effectiveness. *Australian Journal of Political Science*, **45**(2), 245–60.

Bratberg, E., Tjøtta, S., and Øines, T. (2005). Do voluntary international environmental agreements work? *Journal of Environmental Economics and Management*, **50**(3), 583 – 597.

Breitmeier, H., Young, O. R., and Zürn, M. (2006). *Analyzing International Environmental Regimes. From Case Study to Database.* Cambridge, MIT Press.

Breitmeier, H., Underdal, A., and Young, O. R. (2011). The effectiveness of international environmental regimes: Comparing and contrasting findings from quantitative research1. *International Studies Review*, **13**, 1–27.

Downs, G. W. (2000). Constructing effective environmental regimes. *Annual Review of Political Science*, **3**(1), 25–42.

EMEP (2008a). Inventory review. emission data reported under the lrtap convention and the nec directive. status of gridded data. Technical report, CEIP.

EMEP (2008b). Transboundary acidification, euthrophication and ground level ozone in europe in transboundary acidification, euthrophication and ground level ozone in europe in 2006. Status report 2008, EMEP.

Finus, M. and Tjøtta, S. (2003). The oslo protocol on sulfur reduction: the great leap forward? *Journal of Public Economics*, **87**(9-10), 2031 – 2048.

Haas, P. M., Keohane, R. O., and Levy, M. A., editors (1993). *Institutions for the Earth*. Cambridge, MIT Press.

Heckman, J. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. Working Paper 2861, National Bureau of Economic Research.

Helm, C. and Sprinz, D. (2000). Measuring the effectiveness of international environmental regimes. *Journal of Conflict Resolution*, **44**(5), 630–652.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81**(396), 945–60.

Hovi, J., Sprinz, D. F., and Underdal, A. (2003). The oslo-potsdam solution to measuring regime effectiveness: Critique, response, and the road ahead. *Global Environmental Politics*, **3**(3), 74–96.

Keohane, R. (1984). *After Hegemony: Cooperation and Discord in the World Political Economy (Princeton Classic Editions)*. Princeton University Press.

Levy, M. A. (1993). European acid rain: The power of tote board diplomacy. In P. M. Haas, R. O. Keohane, and M. A. Levy, editors, *Institutions for the Earth*, pages 75–133. Cambridge, MIT Press.

Maler, K. G. (1989). The acid rain game. In H. Folmer and E. van Ierland, editors, *Valuation Methods and Policy Making in Environmental Economics*. Elsevier Science Publishers.

Maler, K.-G. and Zeeuw, A. D. (1998). The acid rain differential game. *Environmental and Resource Economics*, **12**(2), 167–184.

Miles, E. L., Underdal, A., Andresen, S., Wettestad, J., Skjærseth, J. B., and Carlin, E. M., editors (2002). *Environmental Regime Effectiveness: Confronting Theory with Evidence*. Cambridge, MIT Press.

Mitchell, R. B. (2002). A quantitative approach to evaluating international environmental regimes. *Global Environmental Politics*, **2**(4), 58–83.

Mitchell, R. B. (2006). Part two: The effectiveness of international environmental regimes - Problem structure, institutional design, and the relative effectiveness of international environmental agreements. *Global Environmental Politics*, **6**(3), 72– 89.

Mitchell, R. B. (2008). Evaluating the performance of environmental institutions: What to evaluate and how to evaluate it? In O. Young, L. King, and H. Schroeder, editors, *Institutions and Environmental Change*, pages 79–114. Cambridge, MIT Press.

Mitchell, R. B. and Deane, G. (2009). Comparing institutional influence: The relative effectiveness of three environmental agreements. Working Paper.

Munton, D., Soroos, M., Nikitina, E., and Levy, M. A. (1999). Acid rain in europe and north america. In O. R. Young, editor, *The Effectiveness of International Environmental Regimes*, pages 155–249. Cambridge, MIT Press.

Murdoch, J. C. and Sandler, T. (1997). The voluntary provision of a pure public good: The case of reduced cfc emissions and the montreal protocol. *Journal of Public Economics*, **63**(3), 331–349.

Murdoch, J. C., Sandler, T., and Sargent, K. (1997). A tale of two collectives: Sulphur versus nitrogen oxides emission reduction in europe. *Economica*, **64**(254), 281–301.

Odén, S. (1968). The acidification of air and precipitation and its consequences on the natural environment. *Swedish Nat. Sci. Res. Council, Ecology Committee*, **1**.

Ringquist, E. J. and Kostadinova, T. (2005). Assessing the effectiveness of international environmental agreements: The case of the 1985 helsinki protocol. *American Journal of Political Science*, **49**(1), 86–102.

Rubin, D. (1974). Estimating the causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**(688-701).

Rubin, D. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics*, **2**, 1–26.

Sandnes, H. (1993). Calculated budges for airborne acidifying components in europe, 1985, 1987, 1988, 1989, 1990, 1991, and 1992. Technical report, Norwegian Meteorological Institute.

Simmons, B. and Hopkins, D. (2005). The constraining power of international treaties: Theory and methods. *American Political Science Review*, **99**(04), 623–631.

Stein, J. V. (2005). Do treaties constrain or screen? selection bias and treaty compliance. *The American Political Science Review*, **99**(4), 611–622.

Underdal, A. (2008). Determining the causal significance of institutions: Accomplishments and challenges. In O. Young, L. King, and H. Schroeder, editors, *Institutions and Environmental Change*, pages 49–78. Cambridge, MIT Press.

UNECE (1999). Protocol to the 1979 convention on lrtap to abate acidification, eutrophication and ground-level ozone. Technical report.

Victor, D. G., Raustiala, K., and Skolnikoff, E. B., editors (1998). *The Implementation and Effectiveness of International Environmental Commitments*. Cambridge, MIT Press.

Wettestad, J. (2002a). *Clearing the Air*. London, Ashgate.

Wettestad, J. (2002b). The convention on long-range transboundary air pollution (clrtap). In E. L. Miles, A. Underdal, S. Andresen, J. Wettestad, J. B. Skjaerseth, and E. M. Carlin, editors, *Environmental Regime Effectiveness: Confronting Theory with Evidence*, pages 197–223. Cambridge, MIT Press.

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MIT Press.

Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *The Review of Economics and Statistics*, **87**(2), 385–390.

World-Bank (2010). World development indicators. CD-ROM.

Young, O. R., editor (1999). *The Effectiveness of International Environmental Regimes*. Cambridge, MIT Press.

Young, O. R. (2001). Inferences and indices: Evaluating the effectiveness of international environmental regimes. *Global Environmental Politics*, **1**(1), 99–121.

Young, O. R. (2003). Determining regime effectiveness: A commentary on the oslo-potsdam solution. *Global Environmental Politics*, **3**(3), 97–104.