

# **QUALITY UNKNOWN TECHNICAL APPENDIXES**

# CONTENTS

<b>APPENDIX A: DATA APPENDIX</b>	<b>3</b>
From the Water: The Problematic Paucity of Existing Data	3
Remote Sensing Data	6
From the Machine: Predictions of Water Quality Using Machine Learning	8
Data and Additional Methodological Points	10
Synthesis of Results	11
Concluding Remarks	15
Notes	16
References	16
<b>APPENDIX B: CAPTURING ENVIRONMENTAL AMENITY VALUES</b>	<b>18</b>
Notes	22
References	22
<b>APPENDIX C: TECHNICAL APPENDIX OF RESULTS</b>	<b>23</b>
Chapter 1	23
Chapter 2	31
Chapter 3	42
Chapter 4	44
Chapter 6	47
References	57
Notes	64
References	64

## APPENDIX A:

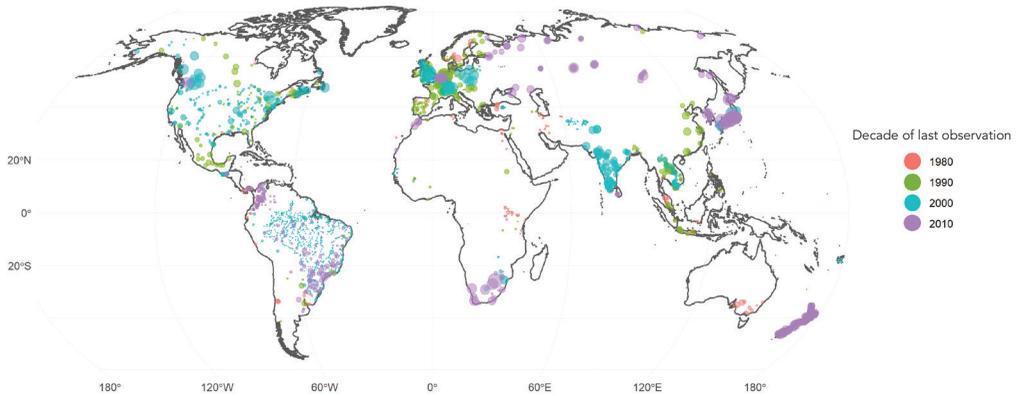
# DATA APPENDIX

What is the global state of water quality? Surprisingly, there is limited scientific consensus on this simple yet crucial policy question. It is known that some pollutants—the classic ones, such as biological oxygen demand (BOD), electrical conductivity (EC), and nitrogen, and the emerging ones, such as microplastics—are present in individual streams, lakes, rivers, or estuaries. Nevertheless, few systematic studies present global evidence on the hotspots and the temporal dynamics of water quality. This paucity of global data on water quality is a major obstacle to understanding impacts and addressing the water quality problem. Recent technological advances and new methods have opened new possibilities. Appendix A presents data collected for this report, as well as two innovative approaches that were developed to build new data sets of water quality.

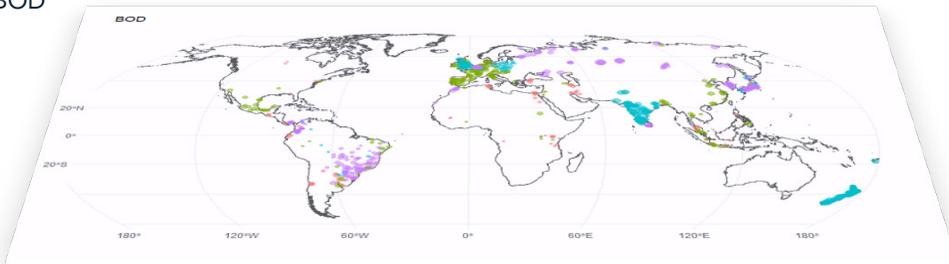
## FROM THE WATER: THE PROBLEMATIC PAUCITY OF EXISTING DATA

GEMStat, which is hosted by the United Nations Environment Programme (UNEP), is the only global repository for water quality data. It contains more than 3.3 million observations of water quality from 2,959 stations in 72 countries and covers 224 parameters, with some observations dating back to the 1960s. Although this database has been vital for the entire water community, it also suffers from significant drawbacks, including the fact that both the data collected and the frequency of collection are highly sporadic, both across and within countries. This is illustrated in map A.1, which shows the location of stations and the date of most recent data available. The size of each dot represents the total number of observations recorded at each station, and the color indicates the decade of the most recent observation. In GEMStat, data collection from many stations ended more than 20 years ago, and large parts of the globe are excluded entirely. These limitations compromise achievement of the Sustainable Development Goal (SDG) 6 on clean water (and sanitation), specifically measurement of indicator 6.3.2 on “Proportion of bodies of water with good ambient water quality.” This suggests the need for urgently building capacity for water quality monitoring in developing countries.

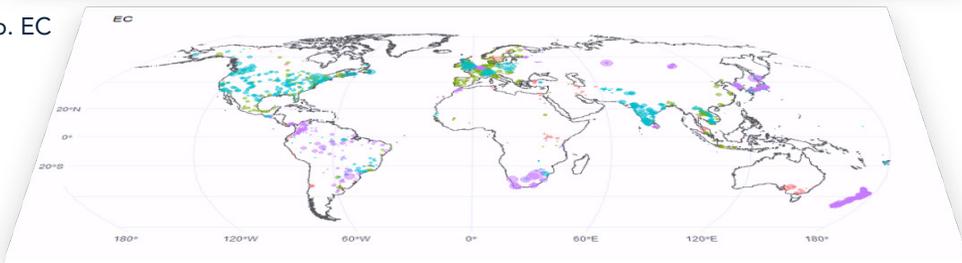
**MAP A.1: GEMStat Data**



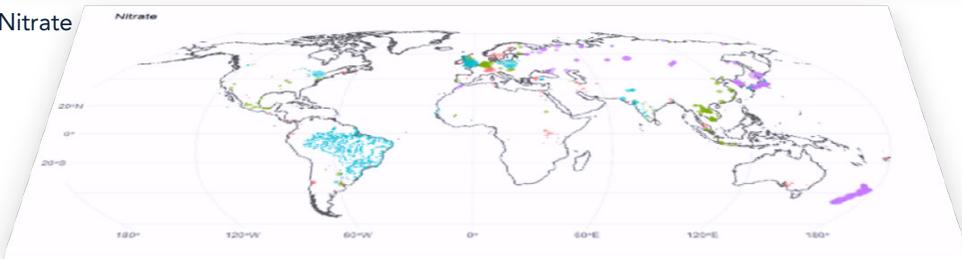
a. BOD



b. EC



c. Nitrate



Source: GEMStat (database), United Nations Environment Programme, Koblenz, Germany, <https://gemstat.org>.  
 Note: BOD = biological oxygen demand; EC = electrical conductivity. The dots in the figure show the distribution of stations in the GEMStat database that record measures of BOD (panel A), EC (panel B), and nitrate (panel C). The top map shows these stations together. The different colors indicate the decade of the most recent observation recorded. Dots vary in size with the total number of observations per station.

Although the GEMStat data set is not exhaustive, many countries, states, regions, or cities have a system of monitoring stations. Adding these data in a global data set closes some gaps but does not detract from the need for better information and monitoring. For this report, an exhaustive search for data from basins, countries, and regions was performed to gain a better sense of the global state of water quality. These are summarized in table A.1.

In addition to these data sets, there was also an attempt to use new avenues of data measurement to shed light on the current state of water quality. Table A.2 summarizes the data that were generated using remote sensing and machine learning techniques.

**TABLE A.1: Overview of Data Used in This Report**

Country/ region	Database name	Time coverage	Notes
Argentina	Matanza-Riachuelo Basin Authority (Autoridad de Cuenca Matanza Riachuelo [ACUMAR])	2008–17	Surface water data from monitoring stations
Brazil, São Paulo state	Environmental Company of São Paulo State (Companhia Ambiental do Estado de São Paulo [CETESB])	1978–2018	Surface water data from monitoring stations
Chile	Superintendence of Sanitary Services (Superintendencia de Servicios Sanitarios [SISS])	2011–18	Municipal drinking water quality data
China	China National Environmental Monitoring Center (CNEMC)	2006–18	Surface water data from monitoring stations
China, Shenzhen	Shanghai Qingyue Environmental Protection Center	2012–14	Tap water quality data are released by water supply companies in China. The historical data are collected by Shanghai Qingyue Organization and are available at <a href="http://data.epmap.org/drink_waters">http://data.epmap.org/drink_waters</a> .
European Union	European Environment Agency database on the status and quality of Europe's rivers (Waterbase)	1965–2012	River water monitoring stations in 39 European countries
Global	GEMStat database	1965–2016	Monitoring station data for predominantly surface river water but also some groundwater and lakes
India	Central Pollution Control Board	1986–2016	Surface water data from monitoring stations
India	Central Water Commission	1963–2017	Surface water data from monitoring stations
India	Central Ground Water Board (CGWB)	2005–09	Groundwater data from monitoring wells
Mekong River basin	Mekong River Commission (MRC)	1985–2014	Surface water data for Cambodia, Lao People's Democratic Republic, Thailand, and Vietnam
Mexico	National Water Information System of National Water Commission (Comisión Nacional del Agua [CONAGUA]) data set	2006–17	Surface water data from monitoring stations for Mexico. Fecal data only available for 2012–17.

**TABLE A.2: Overview of Data Developed in This Report**

Region	Database name	Time coverage	Notes
Global	Remote sensing water quality in lakes (chlorophyll, turbidity, total suspended solids, cyanobacteria, colored dissolved organic matter, floating vegetation, temperature)	2001–12	Remotely sensed data available as monthly lake averages or raster files with a 300-meter resolution. Data are for 421 of the world's largest lakes, including oversampling in Argentina, India, and Mexico.
Global	Machine learning-generated data sets (biological oxygen demand, electrical conductivity, nitrogen)	1992–2010	Global gridded data sets (0.5-degree grids) generated from a model using a machine learning algorithm

## REMOTE SENSING DATA

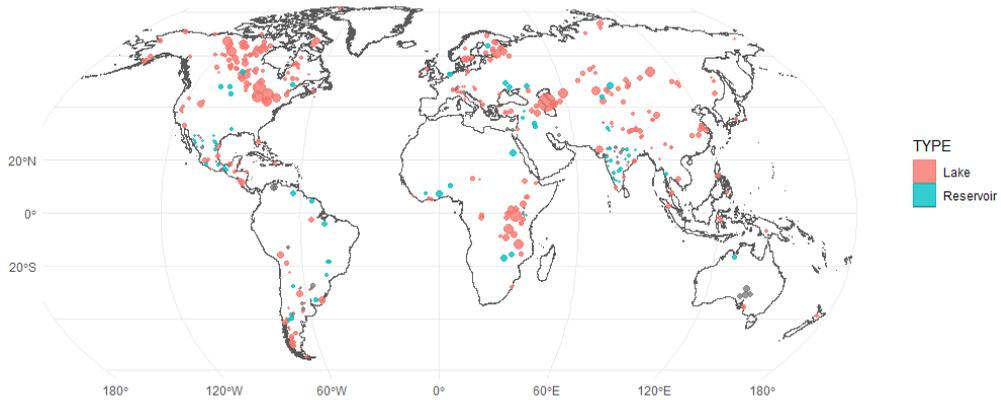
Remote sensing has been a game changer for monitoring environmental indicators, especially air pollution and deforestation. An example is the Hansen dataset that tracks tree cover losses annually (Hansen et al. 2013). For published research alone, the Hansen data have been used or cited by more than 2,500 analyses in less than five years. The database has become one of the most influential scientific contributions, profoundly changing knowledge of deforestation and transforming policies aimed at tackling it. In the same way, remote sensing today could transform the water sector. World Bank projects have tested the possibilities brought by remote sensing in several key bodies of water—for example, in Mexico and Lake Chad. Building on this experience, this report brings remote sensing to scale.

For this report, the largest historical data set of water quality in lakes was developed. It maps eight water quality parameters in 447 lakes and reservoirs around the globe monthly between 2002 and 2012 at a resolution of 300 meters (see map A.2). The data are derived from Envisat MERIS data produced by the European Space Agency. This work, built on the Diversity II experience (Odermatt et al. 2018), extends existing knowledge of water quality by (a) providing the largest global remote sensing database on water quality in lakes and (b) relying on an internally consistent approach, making results comparable across time and space.

More than 350 lakes and reservoirs were selected based on their biodiversity relevance, size, and geographic distribution across the world, and 80 additional lakes were oversampled in three countries where water quality is a prominent issue: Argentina, India, and Mexico. Criteria for selection into the database include size of lake and deemed international importance, based on the Ramsar Convention for wetland conservation established in 1971 by the United Nations Educational, Scientific and Cultural Organization (UNESCO), or LakeNet, a global network of people and organizations dedicated to the conservation of lake ecosystems around the world.<sup>1</sup>

Remote sensing satellites track seven water quality indicators that have a color signal: turbidity, chlorophyll *a* (a proxy parameter for phytoplankton biomasses and the productivity of a lake), colored dissolved organic matter (CDOM), cyanobacteria, lake surface water temperature (LSWT), total suspended matter (TSM), and floating vegetation. More detail of these parameters are given in table A.3.

**MAP A.2: Global Lakes Database, World Bank**



Source: Global Lakes Database, World Bank, Washington, DC

Note: Dots sizes are a function of the total area of lakes derived from the Diversity II lakes database.

**TABLE A.3: Parameters Generated in Remote Sensing Water Quality Data Set**

Parameter	Units	Definition
Chlorophyll <i>a</i>	milligrams per square meter	A photosynthetic pigment that enables green plants to undertake photosynthesis. Its presence in surface water serves as an indicator of an abundance of green vegetation, mainly phytoplankton or algae.
Colored dissolved organic matter	CDOM per square meter	The optically measurable component of dissolved organic matter, made up of a mixture of organic acids from decomposition organic matter.
Cyanobacteria (floating and immersed)	Portion of lake covered in floating or immersed cyanobacteria	Photosynthetic bacteria that share some properties with algae in that they possess chlorophyll and release oxygen during photosynthesis.
Floating vegetation	Portion of lake covered in floating vegetation	Vegetation floating on the lake surface.
Lake surface water temperature	kelvin	Temperature of the top layer of water.

**TABLE A.3: Parameters Generated in Remote Sensing Water Quality Data Set**  
*continued*

Parameter	Units	Definition
Total suspended matter	grams per square meter	TSM in freshwater systems includes chlorophyll, organic influents, and sediment (Cai et al. 2012). High levels of TSM can alter the turbidity characteristics of water (Mallin, Johnson, and Ensign 2009).
Turbidity	FTU	Similar to TSM, turbidity is a measure of cloudiness of water. In scientific terms, turbidity describes the amount of light scattered or blocked by suspended particles. High levels of turbidity are often caused by sediment, organic matter, and—importantly—algae, cyanobacteria, and other phytoplankton.

Note: CDOM = colored dissolved organic matter; FTU = formazin turbidity unit; TSM = total suspended matter.

In contrast to data available from monitoring stations, remote sensing can provide globally consistent, locally relevant high-frequency data on water quality at relatively low marginal cost. Remote sensing is thus a promising tool for monitoring water quality. For instance, the seven indicators tracked here are useful to monitor ecosystem health—a prominent policy question because 80 percent of freshwater biodiversity has disappeared since 1970 (WWF 2016). Additionally, remote sensing is also a robust method for measuring algal blooms, a rising threat to water quality.

Still, many water quality indicators that are of primary relevance for policy makers lack a color signal, making it difficult to track them from space. This important limitation applies to several of the parameters used to assess SDG 6.3.2 targets, including dissolved oxygen, EC, nitrogen, pH, and phosphorus. Because fertilizers are a primary cause of chlorophyll *a*, remote sensing is helpful to model the presence of nitrogen or phosphorus. Yet remote sensing will probably remain a complement to monitoring stations data rather than a substitute. Detection methods for common water quality indicators are presented in table A.4.

## FROM THE MACHINE: PREDICTIONS OF WATER QUALITY USING MACHINE LEARNING

For a broad range of parameters, monitoring stations will remain the dominant tool for collecting data on water quality. This section introduces a novel approach that departs from most models developed so far (Hofstra et al. 2019; van Vliet et al. 2013; Voss et al. 2012; Wen, Schoups, and Van De Giesen 2017).<sup>2</sup> It is based on one observation: Although water quality data are scarce, data that measure the drivers of water quality are

**TABLE A.4: Testing Methodology for Selected Parameters**

	Remote sensing	Water sensor	On-site testing	Laboratory testing
Alkalinity			X	X
Ammonium		X		X
BOD				X
Calcium			X	X
CDOM	X			X
Chloride		X		X
Chlorophyll a	X	X		X
COD				X
Coliforms, fecal				X
Coliforms, total				X
Conductivity		X		X
Cyanobacteria	X			
Dissolved oxygen		X		X
Dissolved solids		X		X
<i>E. coli</i>				X
Magnesium		X		X
Nitrate <sup>a</sup>	X	X		X
Nitrite <sup>a</sup>	X	X		X
pH		X	X	X
Phosphorus		X		X
Potassium		X		X
Sodium		X		X
Sulfate		X		X
Temperature	X	X	X	
TSM	X			X
Turbidity	X	X		X

Note: BOD = biological oxygen demand; CDOM = colored dissolved organic matter; COD = chemical oxygen demand; TSM = total suspended matter.

a. Proxies used in remote sensing

not. These data can then be used to recreate missing values of water quality using machine learning predictive algorithms. Similar machine learning and artificial intelligence applications have recently been booming, including population and poverty mapping at precise scales (Jean et al. 2016). This section demonstrates that such tools can also be harnessed for water quality.

Determinants of BOD, EC, and nitrogen are well-documented. Common to these three parameters are factors such as agricultural runoff containing nitrogen fertilizers, livestock, and the discharge of domestic and industrial wastewater. Even though water quality data are sparse in time and space, data

sets of most of the drivers of water quality are available at the global scale and over time. By using algorithms that combine data on these drivers, BOD, EC, and nitrogen are predicted at a 50-kilometer spatial scale and at a monthly level from 1992 to 2010 across the world. In this way, a globally consistent data set of areas at risk of water pollution is generated for this report.

This work primarily relies on algorithms known as Random Forests, a method that seeks to find the combination of factors that explains observed water quality by estimating thousands of decision trees. As opposed to many machine learning algorithms, Random Forests is a transparent method. It relies on little parametrization and can be consequently applied to a broad range of pollutants in a harmonized methodology. The method is developed on the free statistical software R and uses primarily open-access data, which can easily be adapted to other pollutants and other time periods, depending on the availability of the data. The full list of covariates included in the model and additional methodological points are presented in the following section.

## DATA AND ADDITIONAL METHODOLOGICAL POINTS

The Random Forests model was implemented in R using the caret and ranger libraries. For each pollutant, the data were separated into a training set (80 percent of the observations) and a testing set (20 percent of the observations). Standard practices in the literature were followed: Models were run using 1,000 trees, and the number of parameters included in each tree equaled the square root of the total number of variables.

The variables included in the models are:

- Sanitation provision at the country level from the World Health Organization (WHO)/United Nations Children’s Fund (UNICEF) Joint Monitoring Programme (JMP) for Water Supply, Sanitation and Hygiene, including share of population with at least basic access to sanitation and share of population connected to sewerage (<https://washdata.org/>)
- Annual gridded gross domestic product (GDP) per capita 1990–2015 (Kummu, Taka, and Guillaume 2018)
- Annual population 1990–2015 (Kummu, Taka, and Guillaume 2018)
- Average level of urbanization in 2010 from the Global Rural-Urban Mapping Project, Columbia University (<https://sedac.ciesin.columbia.edu/data/collection/grump-v1/>)
- Annual nitrogen use per hectare in agriculture 1992–2015 at the 0.5-degree scale (Lu and Tian 2017)
- Annual share of croplands and share of forests: annual 1992–2015 data from European Space Agency (<https://www.esa-landcover-cci.org/>)
- Livestock in 2010 (pigs and cattle) from Gridded Livestock of the World – Latest – 2010 (GLW 3) ([https://dataverse.harvard.edu/dataverse/glw\\_3](https://dataverse.harvard.edu/dataverse/glw_3))

- Annual precipitation and temperature 1900–2013 (Willmott and Matsuura 2001)
- Annual runoff data from the Global Water Availability Model (GWAM) 1900–2013
- Elevation and slope
- Harmonized soil database from the Food and Agriculture Organization of the United Nations (FAO)

## SYNTHESIS OF RESULTS

Figure A.1 displays the contribution of the variables in explaining the total variance of the three water quality parameter and ranks the variables by contribution. Maps A.3, A.4, and A.5 are maps of predicted hotspots for EC, nitrogen, and BOD, respectively.

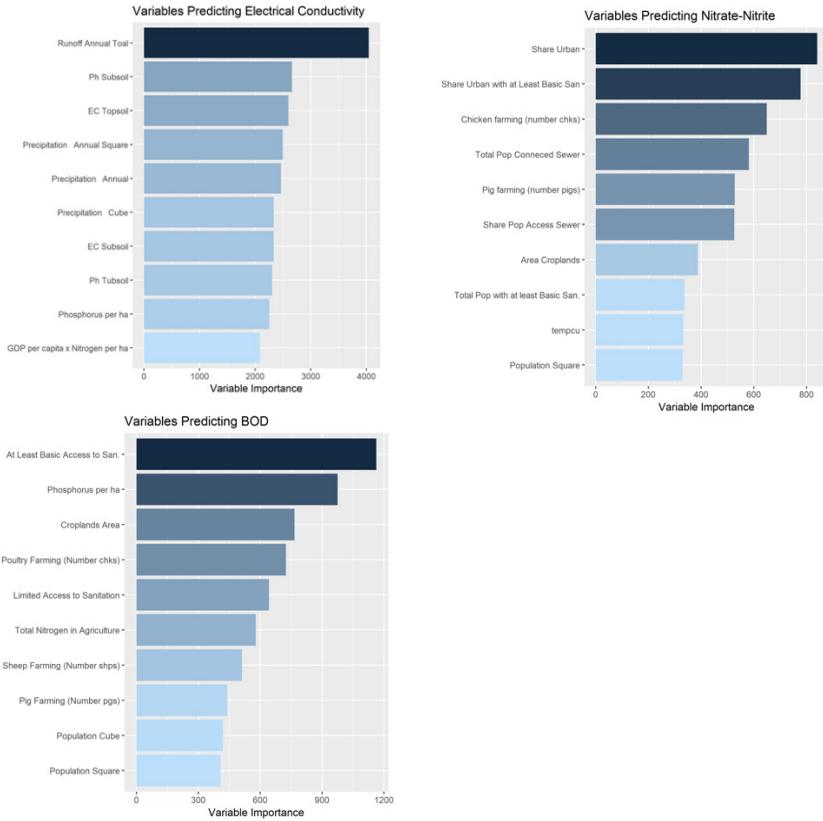
Figure A.1 shows that annual runoff explains a large part of the variance in EC, followed closely by soil characteristics and precipitation. Remarkably, the model predicts EC with good precision—the correlation between the predicted and observed values is 88 percent. Where values are poorly predicted, the predicted values are generally lower than observed values. Therefore, the results of the model can be interpreted as conservative predictions of the risk of poor water quality. Simulated data on EC is used in chapter 3 to estimate the share of the world’s agricultural production at risk from salinity.

Levels of nitrogen found in water around the world are highly correlated with population, sanitation, and agriculture variables, corroborating a large body of evidence and the results presented in chapter 4. Combining these parameters together and adding information on weather and geography allows the model to predict 94 percent of the level of nitrogen observed in available water quality data with few prediction errors. Once the optimal model is found, nitrogen is predicted globally, giving valuable information of levels of nitrogen in areas with no prior observation. These data generated for nitrogen are used in chapter 2 to understand the legacy effect of water pollution on infants’ health in Africa.

Predicting BOD is notoriously more difficult than for other parameters, in large part as a result of the broad range of determinants that often interact with one another. In addition, the factors determining BOD degrade in water, making it more volatile than measures of nitrogen or EC. Consequently, previous models of BOD are often compromised by classification errors. Despite these difficulties, the model developed for this report explains 83 percent of the variance observed in the data.

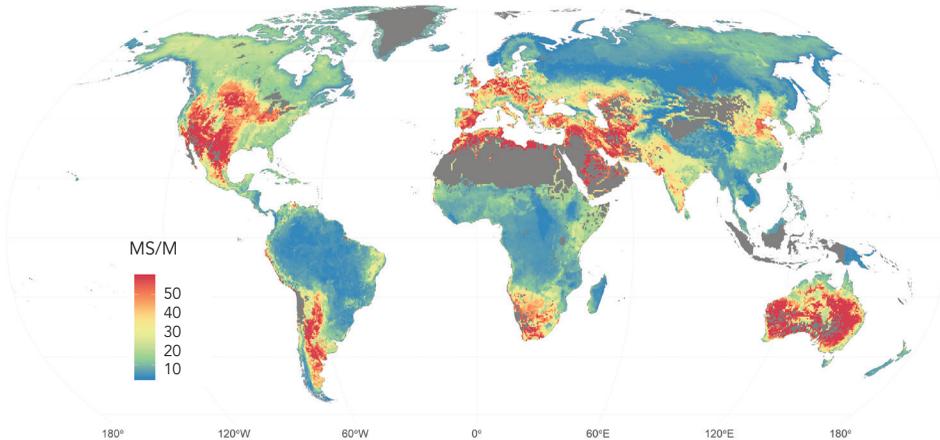
Every model is, by nature, a simplified representation of complex phenomena, and errors are thus inherent to all models. It is arguable that the approach presented in this report has modest errors compared to previous models, yet results must always be taken with some caution.

**FIGURE A.1:** Contribution of Variables in Explaining Total Variance in Electrical Conductivity, Nitrate-Nitrite, and Biological Oxygen Demand



Note: BOD = biological oxygen demand; EC = electrical conductivity; GDP = gross domestic product; ha = hectare.

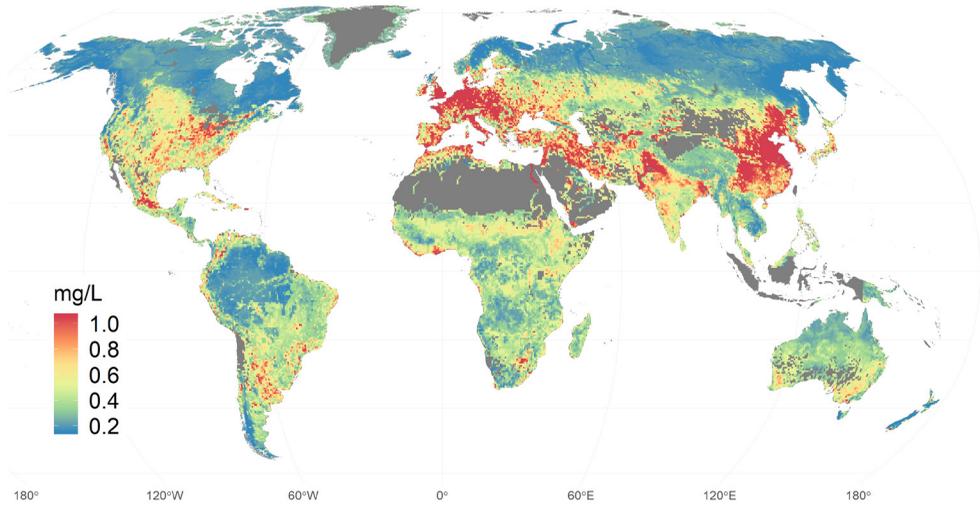
**MAP A.3:** Predicted Hotspots of Electrical Conductivity



Note: Mean = 189.66. EC = electrical conductivity.

---

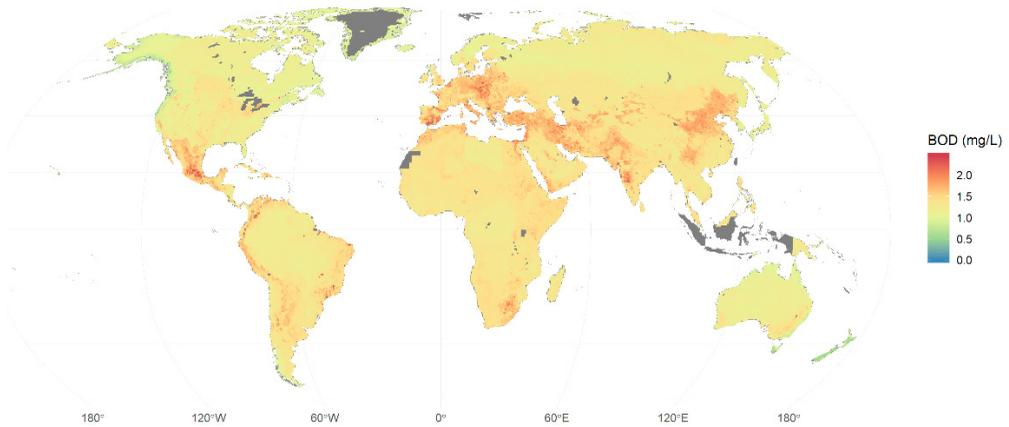
**MAP A.4: Predicted Hotspots of Nitrogen**



Note: Mean = 0.44. NO<sub>x</sub>N = nitrate-nitrite.

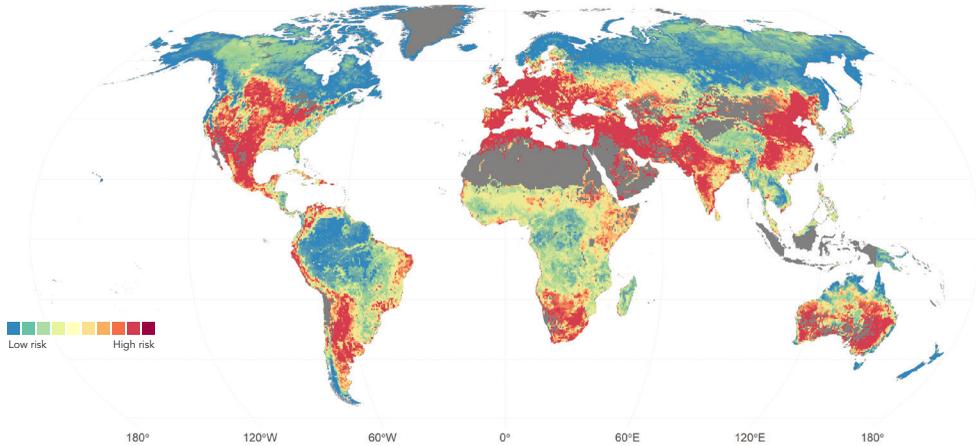
---

**MAP A.5: Predicted Hotspots of Biological Oxygen Demand**



Note: Mean = 1.73. BOD = biological oxygen demand.

## MAP A.6: Water Quality Risk Index



Note: This figure maps a water quality index summarizing global predictions for biological oxygen demand (BOD), electrical conductivity (EC), and nitrogen. Each value is scaled to a common support for comparability, then summed together. Average values for 2000–10 are displayed. Gray areas have no data for one or more parameters.

The findings confirm existing evidence on hotspots and provide new insights. Far from decreasing with economic growth, hotspots of nitrogen pollution are found in agricultural areas of high-income countries, such as the cereal and livestock production areas of Western Europe or the plains of the American Midwest. More importantly, the model highlights the risk of water pollution in vast parts of largely undocumented regions. With complete maps of water quality parameters, it is possible to construct a composite index that gives an overview of the areas most at risk from poor water quality as show in Map A.6.

### SDG 6.3 WATER QUALITY MEASURE

To determine an index of water quality risk, the machine learning data sets on EC, nitrogen, and BOD discussed previously are used. The three variables are “normalized” to be in the same scale and range. This ensures that units of measurement do not determine the weight given to a parameter in the index. The final index is constructed by adding the different parameters, such that a higher value of the index corresponds to poorer water quality.

## CONCLUDING REMARKS

The areas of application of machine learning for water quality monitoring are immense. The method presented in this appendix is flexible and simple enough to model the occurrence of numerous water quality parameters beyond BOD, EC, and nitrogen. In addition, machine learning models offer the possibility of predicting short-term risks of deterioration of water quality with important policy applications. However, all these models are only as good as the existing data used to parametrize the models. Consequently, a degree of uncertainty will generally characterize the results, particularly when data are not obtained using consistent methodologies.

## NOTES

1. For more information on LakeNet, visit the World Lakes website at [www.worldlakes.org](http://www.worldlakes.org).
2. Following an approach similar to the one developed here, many previous models use data on the drivers of water quality to model it. Rather than infer the global link between observed water quality and its drivers from the data, it uses dose–response functions previously estimated in the literature in certain areas of the world. These models have the value added of being hydrologically more correct and using precise local data. The value added of the approach developed here is to infer more complex relationships involving more drivers that can interact together. It also adopts a different philosophy in the sense that it fully lets the data speak. For these reasons, the new model presented here is complementary to the rest of the literature.

## REFERENCES

- Cai, L., G. Zhu, M. Zhu, H. Xu, and B. Qin. 2012. “Effects of Temperature and Nutrients on Phytoplankton Biomass during Bloom Seasons in Taihu Lake.” *Water Science and Engineering* 5 (4): 361–74. doi:10.3882/j.issn.1674-2370.2012.04.001.
- Hansen, M.C., P.V. Potapov, R. Moore, M. Hancher, S.A.A Turubanova, A. Tyukavina, D. Thau, S.V. Stehman, S.J. Goetz, T.R. Loveland, and A. Kommareddy. 2013. “High-resolution global maps of 21st-century forest cover change”. *Science* 342 (6160): 850-853.
- Hofstra, N., C. Kroeze, M. Flörke, and M. T. H. van Vliet. 2019. “Editorial Overview: Water Quality: A New Challenge for Global Scale Model Development and Application.” *Current Opinion in Environmental Sustainability* 36: A1–A5.
- Jean, N., M. Burke, M. Xie, W.M. Davis, D.B. Lobell, and S. Ermon. 2016. “Combining satellite imagery and machine learning to predict poverty.” *Science* 353(6301): 790-94.
- Kummu, M., M. Taka, and J. H. Guillaume. 2018. “Gridded Global Datasets for Gross Domestic Product and Human Development Index over 1990–2015.” *Scientific Data* 6 (5): 180004.
- Lu, C., and H. Tian. 2017. “Global Nitrogen and Phosphorus Fertilizer Use for Agriculture Production in the Past Half Century: Shifted Hot Spots and Nutrient Imbalance.” *Earth System Science Data* 9 (1): 181–92.
- Mallin, M. A., V. L. Johnson, and S. H. Ensign. 2009. “Comparative Impacts of Stormwater Runoff on Water Quality of an Urban, a Suburban, and a Rural Stream.” *Environmental Monitoring and Assessment* 159 (1–4): 475–91. doi:10.1007/s10661-008-0644-4.
- Odermatt, D., O. Danne, P. Philipson, and C. Brockmann. 2018. “Diversity II Water Quality Parameters from ENVISAT (2002–2012): A New Global Information Source for Lakes.” *Earth System Science Data* 10: 1527–49.

- van Vliet, M. T. H., W. H. P. Franssen, J. R. Yearsley, F. Ludwig, I. Haddeland, D. P. Lettenmaier, and P. Kabatad. 2013. "Global River Discharge and Water Temperature under Climate Change." *Global Environmental Change* 23 (2): 450–64.
- Voss, A., J. Alcamo, I. Barlund, F. Voss, E. Kynast, R. Williams, and O. Malve. 2012. "Continental Scale Modelling of In-stream River Water Quality: A Report on Methodology, Test Runs, and Scenario Application." *Hydrological Processes* 26 (16): 2370–84.
- Wen, Y., G. Schoups, and N. Van De Giesen. 2017. "Organic Pollution of Rivers: Combined Threats of Urbanization, Livestock Farming and Global Climate Change." *Scientific Reports* 7: 43289. doi:10.1038/srep43289.
- Willmott, C. J. and K. Matsuura. 2001. "Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1900 - 2014)". [http://climate.geog.udel.edu/~climate/html\\_pages/download.html](http://climate.geog.udel.edu/~climate/html_pages/download.html)
- WWF (World Wildlife Fund). 2016. *Living Planet Report 2016*. Washington, DC: WWF.

## APPENDIX B:

# CAPTURING ENVIRONMENTAL AMENITY VALUES

Even when water is not used for drinking, agriculture, or commercial activity, it still has an economic value. People enjoy living near pristine bodies of water. It provides them with recreation, exercise, and sometimes simply stunning views. However, when the water becomes polluted (giving off odors), becomes clogged with trash, or is too toxic to swim in, the economic value of that body of water can turn negative. Appendix B attempts to determine whether water quality impairs the development and prosperity of urban centers in ways that extend beyond the impacts on health and labor productivity explored in chapter 2. Land values and property prices mirror the wealth, prosperity, and economic activity of the regions in which they are located. More generally, the price of a house is determined by a suite of often interconnected factors such as its features (size, number of rooms, and the quality of construction), its location (access to centers of employment and transport hubs), as well as wider environmental and neighborhood attributes, such as crime, noise, congestion, and public amenities. A house located in more aesthetically appealing surroundings will, therefore, command a higher price than a comparable house in a less desirable neighborhood. The difference in house prices that results is termed the *hedonic* (or implicit) price and provides a measure of the value that consumers place on better environmental amenities of the neighborhood. Because land is immobile, the (desirable and disagreeable) attributes of the surrounding area are reflected (that is, capitalized) in land and property prices. In the context of water quality, this implies that a house by a clean river, offering attractive views or recreational amenities, would sell for a higher price than an equivalent property located close to an unhygienic sewer. An assessment of house price differences, as a result of water quality differences, provides an indication of the demand for cleaner water in the neighborhood. This approach has often been used to infer the value that citizens place on the quality of water in rivers and lakes in their neighborhoods. Indeed, for regulators such as the U.S. Environmental Protection Agency (EPA), it is one of the main methods for analyzing benefits to water quality improvements (Keiser and Shapiro 2018).

## **BOX B.1: Data for Determining Environmental Amenity Value of Cleaner Water**

Data on house prices are based on the asking price of houses for sale. Although the asking price may differ from the final sale price, the two are correlated and do not differ significantly<sup>2</sup>. The real estate database contains almost 450,000 properties with information on standard dwelling characteristics usually used to estimate hedonic price equations, such as lot size, living space, number of rooms and bathrooms, courtyard, and parking availability. Data on water quality are from the official monitoring stations. Due to data limitations, the focus is on parameters that are common across countries and identified as critical in Sustainable Development Goal (SDG) 6. In order to capture the effect of water quality, the regressions include houses within 2.5 kilometers of the monitoring stations. The final database contains more than 75,000 properties.

As with all methods, there are limitations and caveats to this approach. Many harmful water pollutants are colorless, odorless, and impossible to detect without specialized monitoring equipment. As a result, consumer perceptions of water quality differ from the actual quality of water. Evidence suggests that consumers tend to systematically underestimate the extent of water pollution and consequently undervalue the benefits of cleaner water (Poor et al. 2001). There are further and perhaps even stronger biases that result from what is termed *spatial sorting*. If individual preferences diverge, or if there are differences in the sensitivity of people to odor and pollution, those with strong preferences for a cleaner environment will cluster in “greener” neighborhoods than others. The hedonic prices would then reflect the preferences of a subset of the population, yielding biased estimates. In principle, there are technical solutions to all of these problems; however, in practice, lack of data limits what is achievable.

Notwithstanding these caveats, the hedonic approach has been frequently used in developed country settings, where data are more available to assess the benefits of policies and investments to prevent and remove pollutants. Understanding the monetary contribution of water quality to property prices is useful for estimating the additional property tax revenue that can be obtained from the improvement of water quality and offset some of the costs of water treatment.

This section assesses impacts on house prices in three developing countries in Latin America, for which adequate data on house prices, property attributes, amenities, and pollution have been obtained (see box B.1). The first example comes from the Matanza-Riachuelo basin in Argentina. Once a pristine wetland surrounded by sparsely populated settlements, it is now the receptacle of much untreated waste from about

**TABLE B.1: Elasticities of House Prices**

Country	BOD	COD	<i>E. Coli</i>
Argentina	-0.137***	-0.088***	-0.050**
Brazil	-0.075**	-0.096**	-0.035**
Mexico	-0.069**	-0.081***	-0.034***

Note: Table reports elasticity of housing price to biological oxygen demand (BOD), chemical oxygen demand (COD), and *E. Coli* in nearby rivers for Argentina, Brazil, and Mexico.

\*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

20,000 factories, including heavy industry, petrochemicals, transport, chemicals, and a thermal power station.

The second case is from the state of São Paulo in Brazil, which is one of the main hubs of industrial activity. Manufacturing activities include a large number of polluting industries such as heavy metals, manufacturing, plastics, rubber processing, pulp and paper, chemical, and oil-based industries. As a result, the Tietê River, which meanders over 1,000 kilometers of this region, was once classified as among the most polluted in the world.<sup>1</sup> However, its water quality has improved significantly as a consequence of large investments.

The final example contains the most comprehensive coverage with data on house prices and water quality from across Mexico. It captures areas of intensive industrialization and urban activity such as Mexico City and Monterey, as well as smaller and cleaner towns such as La Colina and Puerto Vallarta.

The results reported here focus on the subset of Sustainable Development Goals (SDGs) 6.2 and 6.3 pollutants—biological oxygen demand (BOD), chemical oxygen demand (COD), and *E. Coli*—for which data are available across all three countries. The core results of the analysis are summarized in table B.1. In all three cases, the results consistently show that declining water quality leads to diminishing property prices. Although this may be predictable, the fact that the impacts (elasticities) are so similar across the three disparate economies is surprising.

As an example, table B.1 indicates that a 100 percent reduction in BOD increases house prices by between 6.9 percent in Mexico and 13.7 percent in Argentina. Likewise, a 100 percent decline in COD raises house prices from 8.1 percent in Mexico to 9.6 percent in Brazil, and for *E. Coli*, a 100 percent reduction induces an increase of between 3.4 percent in Mexico to 5 percent in Argentina. The similarity in the marginal impact of improved water quality between the countries suggests that those affected by water quality in public spaces respond in similar ways to the perceived impacts.

A natural question is how many benefits would accrue were a uniform water quality standard to be met in all three countries. For illustrative

**TABLE B.2: Benefits of Improving Water Quality**

		Increase in property values (%)	Average benefit per household (\$)
BOD	Argentina	6.0	13,265
	Brazil	5.6	13,361
	Mexico	5.3	10,158
COD	Argentina	3.2	7,020
	Brazil	6.3	15,036
	Mexico	6.0	11,441
<i>E. Coli</i>	Argentina	4.2	9,263
	Brazil	3.4	8,245
	Mexico	3.1	5,870

Note: Table shows the benefits to improving water quality to the “acceptable” standard prescribed by Comisión Nacional del Agua (CONAGUA). BOD = biological oxygen demand; COD = chemical oxygen demand.

purposes, the standards defined by the Mexican regulator National Water Information System of National Water Commission (Comisión Nacional del Agua [CONAGUA])<sup>3</sup> are used for the simulations reported in table B.2.

The results suggest that there is a high and uniformly robust willingness to pay for better water quality in all three countries. The total benefits accruing to each vary depending on the baseline level of pollution, as well as the number of people benefiting from the improvement.

The need for addressing urban water quality problems is hidden in plain sight. Because those who live near polluted waterways tend to be poor, the investments in water quality improvements would generate a triple win for poverty reduction, urban growth, and the environment.

## NOTES

1. For more information, see [https://pt.wikipedia.org/wiki/Rio\\_Tiet%C3%AA](https://pt.wikipedia.org/wiki/Rio_Tiet%C3%AA) and [http://www.dae.sp.gov.br/index.php?option=com\\_content&id=793:historico-do-rio-tiete&Itemid=53](http://www.dae.sp.gov.br/index.php?option=com_content&id=793:historico-do-rio-tiete&Itemid=53).
2. Even in more sophisticated housing markets, such as in the United States, where buyers and sellers do engage in more negotiations, a significant proportion of houses are sold at their list price. For instance, Case and Shiller (2003) report that 48.4 percent of houses in four major U.S. cities were sold for their asking price.
3. Water quality thresholds provided by CONAGUA:

Parameter	Acceptable	Good	Excellent
BOD (mg/L)	30	6	3
COD (mg/L)	40	20	10
TSS (mg/L)	150	75	25
<b>E. Coli</b> (millions)	0.001	0.0002	0.0001

Note: BOD = biological oxygen demand; COD = chemical oxygen demand; TSS = total suspended solids

## REFERENCES

- Case, K., and R. J. Shiller. 2003. "Is There a Bubble in the Housing Market?" *Brookings Papers on Economic Activity* 2003 (2): 299–362.
- Keiser, D. A., and J. S. Shapiro. 2018. "Burning Waters to Crystal Springs? U.S. Water Pollution Regulation Over the Last Half Century." Economics Working Papers 18016, Iowa State University, Ames, Iowa.
- Poor, J., K. Boyle, L. Taylor, and R. Bouchard. 2001. "Objective versus Subjective Measures of Water Clarity in Hedonic Property Pricing Value Models." *Land Economics* 77 (4): 482–93.

## APPENDIX C:

# TECHNICAL APPENDIX OF RESULTS

This section describes the main empirical and identification techniques used in the report. Following the description, regression results that form the basis for those described in the report are displayed. This section is meant only to facilitate easy access to the main results and will not be printed in the final documents.

## CHAPTER 1

### HOW MUCH DOES IT COST?

Empirical strategy:

For each 0.5-degree grid cell and year between 1991 and 2014, the annual growth rate of the cell's gross domestic product (GDP) is computed adjusted for inflation ( $g$ ). A fixed-effect panel regression is used to determine if  $g$  is impacted by poor water quality measured by biological oxygen demand (BOD). We estimate the following model:

$$GDP\ growth_{i,t-1}^d = \beta_1 polluted_{i,t}^{up} + \beta_2 rainfall_{i,t}^d + \beta_3 rainfall_{i,t}^{d^2} + \beta_4 temp_{i,t}^d + \beta_5 temp_{i,t}^{d^2} + \beta_6 \log\ population_{i,t} + \gamma_i^d + \phi_t + \phi'_{m,t} + \phi''_{m,t-1} + \mu_{c,t} + \epsilon_{i,t} \quad (C.1)$$

Cells are indexed by  $i$  and year by  $t$ . Following the scientific literature, rivers are considered unpolluted when BOD is less than 2 mg/L, moderately polluted when BOD is between 2 and 8 mg/L, and heavily polluted when BOD is more than 8 mg/L. Accordingly, we construct a discrete variable for BOD with three levels: nonpolluted observations (BOD is less than 2 mg/L), moderately polluted observations (BOD is between 2 and 8 mg/L), and heavily polluted observations (BOD is more than 8 mg/L). Sixty percent of the observations have a BOD less than 2 mg/L, and 5 percent of the observations have a value larger than 8 mg/L. We control for rainfall and temperature because both are correlated with water quality and economic activity. For these two weather variables, we allow a quadratic relationship

between them and GDP. Time invariant factors that affect cells' growth rates are controlled by grid cell fixed effects ( $\gamma_i$ ). Global events are captured by year fixed effects ( $\phi_t$ ). Unobserved country changes are accounted for by country-specific time trends ( $\mu_{c \times t}$ ). BOD is measured in each station and year in a different month of a year. To account for seasonal variations in BOD, we include fixed effects for the month during which BOD is measured in the current and the previous years. Standard errors are clustered at the cell level.

Deviations in BOD are likely to be randomly distributed and unexpected and, therefore, orthogonal to any possible confounders. The rich set of fixed effects in this specification isolates localized fluctuations in BOD, facilitating causal inference. The identification strategy, therefore, relies on exogenous changes in water quality upstream. More information on the hydrological link between upstream and downstream regions in the report is detailed in Chapter 1.

## RESULTS

Our main results are presented in table C.1. Column 1 displays the results of the baseline model in equation (C.1). Columns 2 and 3 are subsample analyses for middle-income countries and high-income countries, respectively. Our results suggest a consistent negative impact of poor water quality on GDP growth. In columns 4 through 6, we condense the “polluted” categories into a single bin containing all observations with a BOD larger than 8 mg/L, a threshold beyond which water is heavily polluted. The results remain qualitatively unchanged.

To further corroborate these results, several robustness checks are presented. In table C.2, we use a continuous measure of BOD instead of bins. In columns 1 through 3, we test for a linear relationship between GDP growth and BOD. In columns 4 through 6, we allow a quadratic relationship. The results show that an additional increase in BOD concentrations decreases economic growth.

In table C.3, we test for a broad range of alternative specifications that are usual in the environmental econometrics literature. First, we drop month fixed effects to replicate classic models with only yearly data (column 2). Second, we include a country-specific quadratic time trend instead of a country-specific linear time trend to allow for more complex dynamics (column 3). Third, we exclude major oil-producing countries (Russia and the U.S.), for which economic production is expected to be significantly less affected by water quality (column 4). Fourth, we include an income category-year fixed effects in addition or instead of year fixed effects (columns 5 and 6). Fifth, we eliminate year fixed effects (column 7). Seventh, we account for one lag GDP growth to model convergence as in a Solow growth model. Results are robust to all these alternate specifications. In table C.4, we also replace total GDP by GDP per capita and find qualitatively similar results.

**TABLE C.1: Impact of BOD on GDP Growth**

Variables	1 All countries	2 Middle- income	3 High- income	4 All countries	5 Middle- income	6 High- income
Moderately polluted (vs. nonpolluted)	-1.438*** (0.074)	-1.759*** (0.091)	0.280*** (0.037)			
Heavily polluted (vs. nonpolluted)	-1.980*** (0.100)	-2.509*** (0.138)	-0.012 (0.086)			
Heavily polluted (vs. BOD < 8 mg/L)				-0.804*** (0.086)	-1.160*** (0.124)	-0.285*** (0.078)
Rainfall	0.007*** (0.000)	0.009*** (0.000)	-0.001*** (0.000)	0.007*** (0.000)	0.009*** (0.000)	-0.001*** (0.000)
Rainfall squared	-0.000*** (0.000)	-0.000*** (0.000)	-0.000+ (0.000)	-0.000*** (0.000)	-0.000*** (0.000)	-0.000+ (0.000)
Temperature	-0.560*** (0.109)	5.274*** (0.288)	-1.557*** (0.118)	-0.663*** (0.109)	4.742*** (0.280)	-1.568*** (0.118)
Temperature squared	-0.043*** (0.006)	-0.164*** (0.008)	0.040*** (0.006)	-0.041*** (0.006)	-0.154*** (0.007)	0.041*** (0.006)
Log population	0.087 (0.092)	1.249*** (0.250)	-0.450*** (0.042)	0.060 (0.093)	1.152*** (0.252)	-0.451*** (0.042)
Observations	326,448	231,971	94,477	326,448	231,971	94,477
R <sup>2</sup>	0.192	0.198	0.549	0.191	0.197	0.548

Note: Clustered standard errors in parentheses. BOD = biological oxygen demand; GDP = gross domestic product.

\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , + $p < 0.1$ .

In table C.5, we weight observations based on grid cell population to ensure that results are representative of the economy of the countries studied. The results remain unchanged. Lastly, in table C.6, we validate our results by replacing modeled GDP data with nighttime lights data, an independent measure of economic activity. Our results remain qualitatively unchanged, providing further confidence that the results are not an artifact of the GDP data that are used.

**TABLE C.2: Continuous BOD**

Variables	1	2	3	4	5	6
	All countries	Middle– income	High– income	All countries	Middle– income	High– income
BOD	–0.011*** (0.001)	–0.008*** (0.001)	–0.026*** (0.005)	–0.031*** (0.002)	–0.026*** (0.003)	0.026 (0.016)
BOD squared				0.000*** (0.000)	0.000*** (0.000)	–0.002*** (0.000)
Rainfall	0.007*** (0.000)	0.009*** (0.000)	–0.001*** (0.000)	0.007*** (0.000)	0.009*** (0.000)	–0.001*** (0.000)
Rainfall squared	–0.000*** (0.000)	–0.000*** (0.000)	–0.000+ (0.000)	–0.000*** (0.000)	–0.000*** (0.000)	–0.000+ (0.000)
Temperature	–0.663*** (0.109)	4.883*** (0.281)	–1.562*** (0.118)	–0.653*** (0.109)	4.903*** (0.281)	–1.569*** (0.118)
Temperature squared	–0.041*** (0.006)	–0.157*** (0.007)	0.041*** (0.006)	–0.041*** (0.006)	–0.157*** (0.007)	0.041*** (0.006)
Log population	0.064 (0.093)	1.177*** (0.253)	–0.452*** (0.042)	0.061 (0.093)	1.167*** (0.253)	–0.453*** (0.042)
Observations	326,448	231,971	94,477	326,448	231,971	94,477
R <sup>2</sup>	0.191	0.197	0.548	0.191	0.197	0.548

Note: Clustered standard errors in parentheses. BOD = biological oxygen demand.  
\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , + $p < 0.1$ .

**TABLE C.3: Alternative Specifications**

Variables	1 Base	2 No month	3 Quadratic time trend	4 No oil	5 Income category, year fixed effects	6 Income category, year fixed effects, no time trend	7 No year fixed effects	8 Lag GDP growth
Moderately polluted (vs. nonpolluted)	-1.438*** (0.074)	-1.433*** (0.079)	-1.270*** (0.072)	-1.412*** (0.073)	-1.094*** (0.074)	-1.478*** (0.075)	-1.094*** (0.074)	-1.147*** (0.077)
Heavily polluted (vs. nonpolluted)	-1.980*** (0.100)	-1.954*** (0.103)	-2.007*** (0.097)	-1.981*** (0.099)	-2.026*** (0.107)	-2.093*** (0.109)	-2.026*** (0.107)	-1.729*** (0.117)
Rainfall	0.007*** (0.000)	0.007*** (0.000)	0.007*** (0.000)	0.007*** (0.000)	0.004*** (0.000)	0.005*** (0.000)	0.004*** (0.000)	0.007*** (0.000)
Rainfall squared	-0.000*** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)
Temperature	-0.560*** (0.109)	-0.448*** (0.107)	-0.266* (0.111)	-0.778*** (0.121)	0.777*** (0.119)	2.168*** (0.118)	0.777*** (0.119)	0.332* (0.147)
Temperature squared	-0.043*** (0.006)	-0.042*** (0.006)	-0.047*** (0.006)	-0.036*** (0.006)	-0.088*** (0.006)	-0.120*** (0.006)	-0.088*** (0.006)	-0.074*** (0.008)
Log population	0.087 (0.092)	-0.048 (0.093)	0.299** (0.091)	0.173+ (0.093)	0.175+ (0.091)	0.181+ (0.096)	0.175+ (0.091)	0.441*** (0.122)
Observations	326,448	326,448	326,448	320,368	326,448	326,448	326,448	293,165
R <sup>2</sup>	0.192	0.188	0.198	0.189	0.223	0.217	0.223	0.205

Note: Clustered standard errors in parentheses. GDP = gross domestic product.

\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , + $p < 0.1$ .

**TABLE C.4: GDP per Capita Growth**

Variables	1	2	3	4	5	6
	All countries	Middle– income	High– income	All countries	Middle– income	High– income
Moderately polluted (vs. nonpolluted)	–0.701*** (0.084)	–0.948*** (0.119)	–0.255** (0.078)			
Heavily polluted (vs. nonpolluted)				–1.334*** (0.072)	–1.535*** (0.088)	0.312*** (0.034)
Heavily polluted (vs. BOD < 8 mg/L)				–1.796*** (0.098)	–2.143*** (0.133)	0.049 (0.085)
Rainfall	0.007*** (0.000)	0.009*** (0.000)	–0.001*** (0.000)	0.007*** (0.000)	0.010*** (0.000)	–0.001*** (0.000)
Rainfall squared	–0.000*** (0.000)	–0.000*** (0.000)	0.000* (0.000)	–0.000*** (0.000)	–0.000*** (0.000)	0.000* (0.000)
Temperature	–0.711*** (0.098)	3.009*** (0.220)	–0.869*** (0.103)	–0.640*** (0.097)	3.262*** (0.222)	–0.854*** (0.104)
Temperature squared	–0.038*** (0.005)	–0.109*** (0.006)	–0.014* (0.005)	–0.039*** (0.005)	–0.114*** (0.006)	–0.014** (0.005)
Observations	358,816	252,237	106,579	358,816	252,237	106,579
R <sup>2</sup>	0.178	0.185	0.524	0.179	0.186	0.524

Note: Clustered standard errors in parentheses. BOD = biological oxygen demand; GDP = gross domestic product.

\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , + $p < 0.1$ .

**TABLE C.5: Weighted Regressions**

Variables	1	2	3	4	5	6
	All countries	Middle-income	High-income	All countries	Middle-income	High-income
Moderately polluted (vs. nonpolluted)	-0.772*** (0.058)	-0.923*** (0.068)	0.131* (0.059)			
Heavily polluted (vs. nonpolluted)	-1.058*** (0.075)	-1.346*** (0.092)	-0.324** (0.120)			
Heavily polluted (vs. BOD < 8 mg/L)				-0.441*** (0.060)	-0.626*** (0.072)	-0.443*** (0.093)
Rainfall	0.003*** (0.000)	0.003*** (0.000)	-0.001*** (0.000)	0.003*** (0.000)	0.003*** (0.000)	-0.002*** (0.000)
Rainfall squared	-0.000*** (0.000)	-0.000*** (0.000)	0.000+ (0.000)	-0.000*** (0.000)	-0.000*** (0.000)	0.000+ (0.000)
Temperature	-0.375* (0.151)	2.666*** (0.517)	-1.144*** (0.180)	-0.436** (0.154)	2.334*** (0.506)	-1.138*** (0.180)
Temperature squared	-0.004 (0.004)	-0.066*** (0.010)	0.044*** (0.008)	-0.002 (0.004)	-0.058*** (0.009)	0.043*** (0.008)
Log population	-2.434*** (0.398)	-3.032*** (0.494)	0.208 (0.434)	-2.499*** (0.399)	-3.086*** (0.493)	0.215 (0.434)
Observations	326,448	231,971	94,477	326,448	231,971	94,477
R <sup>2</sup>	0.213	0.192	0.514	0.212	0.191	0.514

Note: Clustered standard errors in parentheses. BOD = biological oxygen demand.

\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , + $p < 0.1$ .

**TABLE C.6: Nighttime Light**

Variables	1 All countries	2 Middle-income	3 High-income
Heavily polluted (vs. BOD < 8 mg/L)	-0.050*** (0.003)	-0.034*** (0.003)	-0.072*** (0.006)
Rainfall	-0.000*** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)
Rainfall squared	0.000* (0.000)	0.000 (0.000)	0.000*** (0.000)
Temperature	-0.114*** (0.003)	-0.101*** (0.008)	-0.187*** (0.008)
Temperature squared	0.002*** (0.000)	0.002*** (0.000)	0.004*** (0.000)
Log population	-0.008 (0.007)	0.054*** (0.011)	-0.036** (0.011)
Observations	309,519	218,596	90,923
R <sup>2</sup>	0.353	0.427	0.298

Note: Clustered standard errors in parentheses. BOD = biological oxygen demand.  
\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , + $p < 0.1$ .

## CHAPTER 2

### NITROGEN LEGACY IN INDIA

To estimate the long-run health impacts of childhood nitrate-nitrite exposure, the research design exploits quasi-random variation in exposure to nitrate-nitrite pollution experienced by different birth cohorts in various districts. The focus is on height because it is derived from well-established worldwide protocols for measurement, is commonly used to capture the cumulative effects of health shocks, and is a well-known proxy for overall health and long-term adult well-being. Specifically, the analysis compares height outcomes between exposed and nonexposed cohorts, controlling for average differences in these outcomes across birth years and across districts. The estimating equation (C.2) for individual-level outcome  $Y$  of person  $i$  during time  $t$  and born in district  $d$  and state  $s$  presented below.

Empirical strategy:

$$Y_{idst} = \alpha + \beta \text{frac exposure}_{idst}^{[birth, 3]} + \lambda X_{dst} + \gamma D_{ids}^W + \rho_{mt} + \rho_s t + \mu_d + \epsilon_{idst} \quad (C.2)$$

$\text{frac exposure}_{idst}^{[birth, 3]}$  is the fraction of years from the time of birth to age 3 that a person was exposed to nitrate-nitrite level that exceeds safe limits in their birth district. It serves as a measure of cumulative pollution exposure in early life during generally accepted critical periods for biological growth and development. These values are recorded from upstream districts exploiting the natural flow of rivers and the upstream/downstream relationship of the districts described in Chapter 1. This helps to exploit quasi-random variation in pollution that originates upstream and yet flows downstream to other districts. The analysis then uses these spillovers to ascertain how much of the health impact persists in the next district downstream of pollution incidence.

The analysis compares later-life height among cohorts with more and less nitrate-nitrite exposure, controlling for birth year, birth month, district fixed effects, and state trends. In this way, the analysis exploits within-district variation in birth timing relative to pollution exposure to identify  $\beta$ . The birth year and birth month fixed effects ( $\rho_{mt}$ ) are included to account for age effects in health outcomes as well as unobserved national or seasonal shocks such as macroeconomic conditions or seasonal weather patterns, which might otherwise confound the relationship between pollution exposure and height. Similarly, district fixed effects are included to control for any time-invariant unobservable differences between districts that can affect health. For example, access to local nutrition programs is one such factor that may be constant across individuals born in the same location. The analysis also includes state trends ( $\rho_s t$ ) to flexibly control for heterogeneous changes in demographic factors, technological progress in agriculture, and other policies that differ across states. A number of other district and household specific

variables are included in the analysis:  $X_{dst}$  is a vector of district time-varying variables including temperature and precipitation and concentrations of other water quality indicators like fecal coliform, whereas  $D_{ids}^W$  is a control for household characteristics such as religion and caste, which are salient to the Indian context. Lastly, the analysis uses cluster-robust standard errors to account for within-district clustering of errors and arbitrary correlation of observations across time.

In this way, the relationship between water pollution and height ( $\beta$ ) is identified by removing any confounding differences attributable to location and time following an established empirical method in applied statistics. Furthermore, because the identification strategy uses multiple exposure events over time and space, it alleviates concerns that the results are being driven by confounding factors to health that may be correlated with single-exposure events.

## RESULTS

Table C.7 presents the main results. Column 1 presents results from the preferred specification. Columns 2 through 4 use an indicator of exposure in utero, in the birth year, and at age 1 rather than a cumulative measure. The results show a lowering of height with exposure, but these effects are not always significant compared to the effects from cumulative exposure in column 1. Column 5 includes an indicator for whether concentrations of fecally derived bacteria related to poor sanitation such as fecal coliform are above desired limits to confirm that other correlated water quality indicators are not driving the result. Exposure from nitrate-nitrite continues to be statistically significant, and the magnitude is even higher. This suggests that exposure to nitrate-nitrite matters for health, in addition to exposure from excreta-related bacteria. In column 6, stricter control of birth month by birth year fixed effects are included to control for unobserved factors that are constant across all individuals born in the same year and month. Results are qualitatively similar. So far, the results have focused on the persistence of water quality impacts in downstream areas by measuring the direct spillover externality imposed by upstream pollution. In contrast, column 7 provides the estimated impact of the within-district externality by measuring the impact of pollution on health in the same district. To do so, the impact of the local district-level pollution variable is instrumented with the upstream value. This effectively uses variations in local water quality that are induced by exogenous upstream concentrations. The validity of this approach rests on the assumption that pollution from far away distances affects health—but only through its effect on local pollution concentrations. The approach exploits the fact that the decision to pollute upstream is orthogonal to downstream health inputs and that pollution flows downstream. Diagnostic statistics for instrument relevance such as the Kleibergen-Paap F statistic shows that the instrument is very strong. Because water pollution decays with river flow and time, one would predict that the downstream impact

is smaller in magnitude than the within-district health impact. Indeed, we find exactly that, with instrumental variable point estimates relatively much larger than the corresponding downstream spillover impact in column 1.

In order to examine the possibility that these results are driven by spurious spatial or temporal patterns, the analysis is subjected to falsification tests. The first test involves reestimating the equation while replacing each individual’s exposure condition with exposures that occur for six different four-year periods before or after birth. The resulting coefficient estimates are plotted in figure C.1 against the different window periods of exposure. All the “shifted” coefficients are smaller than the “true” coefficient, plotted at 0-3, and are all statistically insignificant. The second test involves replacing the upstream pollution variable with a falsified value using pollution data from the nearest off-river region farther downstream—a location from where the pollution cannot flow (upstream) to areas where the health outcomes are

**TABLE C.7: Impact of Pollution Exposure on Health, India**

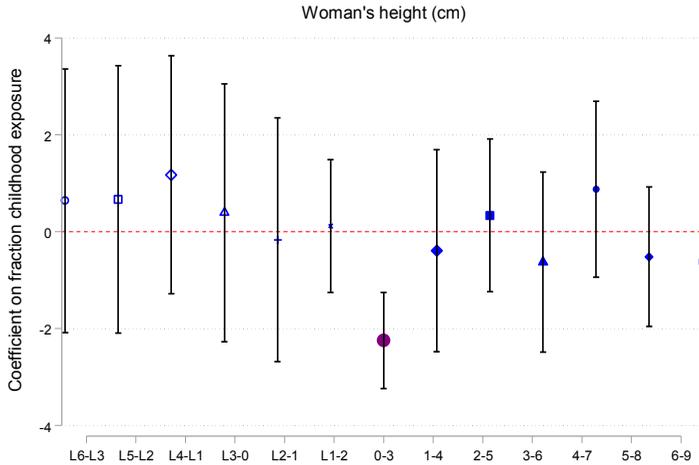
	Dependent variable: height (cm)						
	1	2	3	4	5	6	7
Fraction early childhood N exposure	-2.246*** (0.497)				-3.044*** (0.996)	-1.963*** (0.506)	-2.819*** (0.645)
Exposure in utero		0.541 (0.463)					
Exposure at birth			-0.385 (0.458)				
Exposure at age 1 year				-0.411 (0.392)			
Observations	19138	17399	17618	17417	13862	19138	17755
Mean dependent variable	151.6	151.6	151.6	151.7	151.4	151.6	151.6
R <sup>2</sup>	0.0793	0.0812	0.0795	0.0769	0.0656	0.0908	0.0186
RMSE	6.082	6.076	6.093	6.114	6.089	6.046	5.956
Birth year fixed effects	Y	Y	Y	Y	Y		Y
Birth month fixed effects	Y	Y	Y	Y	Y		Y
District fixed effects	Y	Y	Y	Y	Y	Y	Y
State trends	Y	Y	Y	Y	Y	Y	Y
Weather controls	Y	Y	Y	Y	Y	Y	Y
Fraction early childhood fecal coliform exposure					Y		
Birth year by birth month fixed effects						Y	
Kleibergen–Paap F statistic							20.34 (F = 21.64)

Note: Standard errors in parentheses are clustered at the district level. cm = centimeter; RMSE = root mean square error

\*p < 0.10, \*\*p < 0.05, \*\*\*p < 0.01.

measured. Table C.8 reveals that there is no significant impact of the falsified value on health, suggesting that the upstream variable utilized in the analysis is indeed isolating quasi-random variation in pollution.

**FIGURE C.1:** Different Window Periods of Exposure



Note: Estimated coefficients from variants of the main regression equation, in which the period of pollution exposure is shifted by six four-year periods (horizontal axis) from the main 0-3 period. Each marker's vertical position therefore measures the estimated impact of exposure at the appropriate period of exposure. For example, the purple marker represents the impact of exposure discussed in the report. Other markers represent the impact of "placebo" exposures. Error bars represent 95% confidence intervals. cm = centimeter.

**TABLE C.8:** Impact of Pollution Exposure on Health: Falsification Test

	Placebo districts	
	1	2
Fraction childhood exposure	0.106 (0.430)	-0.010 (0.413)
Observations	23338	23338
R <sup>2</sup>	0.0614	0.0622
RMSE	5.796	5.821
Birth year fixed effects	Y	Y
Birth month fixed effects	Y	Y
District fixed effects	Y	Y
State trends	N	Y
Weather controls	Y	Y

Note: Standard errors in parentheses are clustered at the district level. RMSE = root mean square error  
 \* $p < 0.10$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

## NITROGEN LEGACY IN VIETNAM

A similar empirical approach is used for Vietnam. Instead of focusing on adults like in India, this analysis focuses on children ages 4 to 12 and measures the impact of early childhood exposure on their height-for-age (HAZ) scores (Table C.9).

**TABLE C.9: Impact of Pollution Exposure on Health, Vietnam**

HAZ scores	1	2
Fractional exposure to nitrate–nitrite	−0.776** (0.338)	−0.779** (0.337)
Birth year fixed effects	Y	Y
Birth month fixed effects	Y	Y
Commune fixed effects	Y	Y
Province trends	Y	Y
Other controls	precipitation, temperature, ethnicity (tribe), sex, conductivity, phosphorus, water treatment at home, household asset value, years of education of head, farm vs. nonfarm household	
N	691	691
R <sup>2</sup>	0.132	0.156

Note: Standard errors in parentheses are clustered at the commune level. HAZ = height-for-age.

\* $p < 0.10$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

## NITROGEN LEGACY IN AFRICA

The estimating equation (C.3) for health-level outcome of child  $i$  born in region  $g$  in year  $y$  is presented below.

Empirical strategy:

$$health_{igy} = WQ_{gy} + X_{igy} + W_{gy} + \tau_y + \gamma_y * \mu_c + \theta_g + \varepsilon_{igy} \quad (C.3)$$

$health_{igy}$  is either a dummy indicating whether the child in a household is stunted (HAZ is  $\leq 2$ ) or HAZ scores of children.  $WQ$  is the average nitrate level in the region the household is located during the months of a mother's pregnancy—in utero exposure is considered as the average of the month of birth and the eight previous months.  $X$  is a set of controls that includes household variables: whether it is in a rural location; an indicator for improved sanitary facilities, improved water source, and no sanitation facility (open defecation); child age in months; age of mother when giving birth; whether child is a girl; household wealth index; body mass index (BMI) of mother; an index of mother empowerment (health decisions); and mother's years of education and mother's partner's years of education.  $X$  also includes community variables: percentage of improved water source, improved sanitation, and open defecation and total population of urban

area (in the upstream-downstream analysis).  $W$  is controls for weather (temperature and precipitation).  $\tau_y$  is a year fixed effect.  $\gamma_y * \mu_c$  is a year-specific country trend.  $\theta_g$  is a fixed effect for the region of pollution where the household is located (grid cell). The estimation errors are clustered at this region level (grid cell). Table C.10 shows the results.

**TABLE C.10: Impact of Pollution Exposure on Health, Africa**

In utero exposure	Stunting	HAZ	Stunting	HAZ
Downstream of urban	0.0172*** (0.00636)	(0.0228)		-0.0729***
Downstream of urban x rural			0.0209*** (0.00597)	-0.0848*** (0.0222)
Fixed effects	Year month of birth, grid cell			
Other controls	as indicated in the empirical strategy			
N	204,886	204,886	204,886	204,886
R <sup>2</sup>	0.106	0.143	0.106	0.144

Note: Standard errors in parentheses are clustered at grid-cell level. HAZ = height-for-age. \* $p < 0.10$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

## BOX C.1: ESTIMATING GLOBAL EFFECTS OF NITRATE POLLUTION ON HEALTH

### Empirical Strategy:

A basic cross-country regression of the following form is estimated, where  $c$  indexes countries and  $y$  indexes years:

$$\text{stunting}_{cy} = \beta * \text{NO3}_{cy} + \alpha_c + \gamma_y + X_{cy}\theta + \epsilon_{cy} \quad (\text{C.4})$$

$\text{NO3}_{cy}$  is the mean concentration of nitrate pollution in a country derived from a machine-learning algorithm described in chapter 1 and Appendix A.  $\text{stunting}_{cy}$  is the share of children younger than 5 years that are stunted in a country. The goal is to estimate  $\beta$ , which is the elasticity of stunting to increases of nitrates in the water. To ensure that the result is not an artifact of trending variables, the model controls for year fixed effects ( $\gamma_y$ ). To rule out concerns that nitrate pollution is merely a reflection of general economic development, a control for GDP is included in  $X_{cy}$ . Similarly, to rule out concerns that the resulting impact on stunting is driven by spurious changes in certain countries or by genetic differences, the model accounts for region fixed effects ( $\alpha_c$ ) and controls for average height of mothers of measured children in  $X_{cy}$ , along with other additional time-varying controls such as female literacy rates and share of population with unimproved water and unimproved sanitation—factors that are likely to be correlated with stunting rates. Results are show in Table C.1.1

**TABLE C.1.1: Estimating Global Effects on Nitrate Pollution on Health**

	1	2
	Percent stunting younger than 5 years	
Nitrate (mg/L)	6.0393* (2.793)	4.4382*** (1.164)
ln(GDP)	6.2838 (6.344)	0.0312 (0.687)
Women's height (< 145 cm)	0.093 (0.648)	1.0896*** (0.209)
Women's literacy	-0.1786 (0.129)	0.0859 (0.066)
Unimproved water	0.0469 (0.155)	0.2882*** (0.057)
No toilet	0.2627* (0.13)	0.0038 (0.045)
Year fixed effects	Y	Y
Region fixed effects	Y	Y
Region time trends		Y
N (DHS)	149	126
R <sup>2</sup>	0.436	0.857

Note: Standard errors in parentheses are clustered at country level. cm = centimeter; DHS = Demographic and Health Surveys; GDP = gross domestic product.  
\*p < 0.10, \*\*p < 0.05, \*\*\*p < 0.01.

## BOX C.2 MANY POLLUTANTS, MANY IMPACTS

The estimating equations (C.5-C.7) for measuring the impact of poor water quality on various health-related outcomes are presented below.

Empirical strategy:

### Mexico

$$hosp_{mt} = WQ_{m,y} + WQ_{m,y} * month + total\ hosp_{mt} + X_{m,y} + \tau_t + \gamma_y + \theta_m + \varepsilon_{m,y} \quad (C.5)$$

$$expenditure_{i,y} = WQ_{y,m} + \theta_y + \gamma_m + X_{i,y}\beta + \varepsilon_{i,m,y} \quad (C.6)$$

$hosp_{mt}$  is the logarithm or inverse hyperbolic transformation of the count of diarrhea occurrence in municipality  $m$  and month  $t$ ;  $expenditure_{i,y}$  is the logarithm or inverse hyperbolic transformation of expenditure on health goods or services of household  $i$  in year  $y$  at values of 2014;  $WQ$  is the logarithm or inverse hyperbolic transformation of water quality parameters or an indicator if water quality in upstream municipality  $m$  is above a certain threshold;  $total\ hosp_{mt}$  is the logarithm or inverse hyperbolic transformation of the count of all hospitalization of children up

to age 5 in the same period;  $X_{my}$  is a squared function of precipitation and temperature in municipality  $m$  over the year  $\tau_t$  (data between 2000 and 2014);  $\tau_t$ ,  $\gamma_y$ , and  $\theta_m$  are respectively month, year, and municipality fixed effects. Errors are robust at the municipality level.

$X_{iy}$  is a set of controls that includes quadratic size of household (number of members); share of women and children in the household; educational level of household head; age of household head; group of size of municipality household is in (number of inhabitants); indicator if the household head has health insurance (Seguro Popular); indicator if the household has public supplied water, sewage, and electricity; logarithm of per capita total expenditures (monetary and nonmonetary, as an approach for level of permanent income); value of minimum wage of municipality at values of 2014; and squared precipitation and temperature. Errors are clustered at the municipality level. The estimations are performed separately for three levels of income (based on total expenditure). Results are shown in Table C.2.1

**TABLE C.2.1: Mexico**

	1	2	9	10	11
	ihs (count illnesses in children younger than 1 year)	ihs (count illnesses in children up to 1 year)	ihs (health expenditure) Tercile 1	ihs (health expenditure) Tercile 2	ihs (health expenditure) Tercile 3
BOD > 6 mg/L	0.0749** (0.036)	0.0828* (0.0457)	0.204* (0.0843)	0.109 (0.109)	0.155 (0.181)
Fixed effects	Municipality, year, month				
Other controls	Controls for total admission of child up to 5 years, BOD dummy interacted with month and temperature squared and precipitation squared				
N	19,076	19,076	4750	4644	4444
R <sup>2</sup>	0.031	0.052	0.06	0.06	0.062

Note: Standard errors in parentheses are clustered at municipality level. BOD = biological oxygen demand; ihs = inverse hyperbolic sine.

\*p < 0.10, \*\*p < 0.05, \*\*\*p < 0.01.

**TABLE C.2.2: Brazil**

	1	2	3	4
ihs (count illnesses in children younger than 2 years)				
ihs (conductivity)	0.0681*** (0.0254)			
ihs (ammonia nitrogen)		0.142*** (0.0393)		
ihs (Kjeldahl nitrogen)			0.115** (0.0515)	
ihs (nitrite)				0.121*** (0.0438)
ihs (fecal coliforms)				
Fixed effects		Municipality, year-month		
Other controls	Microregion trends, temperature and precipitation			
N	1,504	1,435	1,477	1,431
R <sup>2</sup>	0.234	0.252	0.242	0.242

Note: Standard errors in parentheses are clustered at municipality level. ihs = inverse hyperbolic sine. \*p < 0.10, \*\*p < 0.05, \*\*\*p < 0.01.

## Brazil

$$disease_{mt} = WQ_{mt} + X_{mt} + \tau_t + \gamma_t * \mu_M + \theta_m + \varepsilon \quad (C.7)$$

$disease_{mt}$  is the logarithm of count of diarrhea or dehydration occurrence in municipality  $m$  and month  $t$ ;  $WQ_{mt}$  is the logarithm of average water quality of a parameter (fecal coliforms, conductivity, ammonia, Kjeldahl nitrogen, and nitrite) over monitoring stations of municipality upstream of  $m$  in month  $t$ ;  $X_{mt}$  is the logarithm of temperature and of precipitation in municipality  $m$  and month  $t$ ;  $\tau_t$  and  $\theta_m$  are time (year-month) and municipality fixed effects; and  $\gamma_t * \mu_M$  is a time (year-month) trend in microregion (group of municipalities)  $M$ . Errors are clustered at the municipality. Results are show in Table C.2.2

## Salt Pollution: A Pinch Too Much?

The estimating equations (C.8-C.10) for measuring the impact of salinity on health-related outcomes are presented below.

### Empirical strategy:

$$ihs(infant\ deaths_{imt}) = \alpha_1 + \alpha_2 * ECz_{imt} + \alpha_3 * Precip_{it} + \alpha_4 * Temp_{it} + \delta_i + \theta_{mt} + \varepsilon_{imt} \quad (C.8)$$

$$ihs(fetal\ deaths_{imt}) = \alpha_1 + \alpha_2 * ECz_{imt} + \alpha_3 * live\ births_{imt} + \alpha_4 * Precip_{it} + \alpha_5 * Temp_{it} + \delta_i + \theta_{mt} + \varepsilon_{imt} \quad (C.9)$$

**TABLE C.2.3: Salt Pollution: A Pinch Too Much?**

	(1)	(2)	(3)
	ihf (infant deaths)	ihf (fetal deaths)	APGAR < 7
Electrical conductivity z-score	0.00267* (0.00149)	0.0401** (0.0170)	0.00234*** (0.000605)
Includes ihf (total births)	N	Y	N
Other controls	Annual precipitation, average temperature		
Year fixed effects	Y	Y	Y
Month fixed effects	Y	Y	Y
Municipality fixed effects	Y	Y	Y
N	77,261	71,218	2,214,042
R <sup>2</sup>	0.680	0.756	0.016

Note: Standard errors in parentheses are clustered at the monitoring station x month x year level. APGAR = Appearance, Pulse, Grimace, Activity, and Respiration; ihf = inverse hyperbolic sine.

\*p < 0.10, \*\*p < 0.05, \*\*\*p < 0.01.

$$APGAR_{it} = \alpha_1 + \alpha_2 * ECz_{imt} + \alpha_3 * precip_{it} + \alpha_4 * temp_{it} + \delta_i + \theta_{mt} + \varepsilon_{imt} \quad (C.10)$$

*infant (fetal) deaths<sub>imt</sub>* are total infant (fetal) deaths in municipality *i*, month *m*, and year *t*; *ECz<sub>imt</sub>* is the z-score of electrical conductivity in the monitoring station upstream to municipality *i* in the nine months prior to *t*. *live births<sub>imt</sub>* are total live births in municipality *i*, month *m*, and year *t*; and *precip<sub>it</sub>* and *temp<sub>it</sub>* are total rainfall and average temperature. *APGAR<sub>i</sub>* is a dummy variable equal to 1 if a child’s APGAR (Appearance, Pulse, Grimace, Activity, and Respiration) score was 6 or below one minute after birth (considered poor health) and 0 otherwise. Identification is predicated on the upstream/downstream relationship of the monitoring station/municipality that is described in Chapter 1. Results are show in in Table C.2.3

## ANNEX CA: THE MANY UNCERTAINTIES OF ARSENIC CONTAMINATION IN DRINKING WATER

The estimating equations (C.11-C.12) for measuring the impact of arsenic on health and labor-related outcomes are presented below.

Empirical strategy:

$$ihf(admissions_{cmyi}) = \alpha_1 + \alpha_2 * ars_{acmy} + \alpha_3 * ihf(hospital_{cmy}) + \delta_p + \omega_{my} + \varepsilon_{cmy} \quad (C.11)$$

$$hours\ worked_{im_y} = \alpha_1 + \alpha_2 * ars_{cmy} + X_i' * \gamma + \delta_{mun} + \omega_{my} + \varepsilon_{im_y} \quad (C.12)$$

*admissions<sub>cmyi</sub>* is the count of admissions in commune *c*, month *m*, year *y*, for ICD-10 category *i*; *ars<sub>acmy</sub>* is an indicator if arsenic is greater than

10 µg/L;  $hospital_{cmy}$  is total hospital admissions for all ICD-10 codes;  $\delta_p$  is regional fixed effects; and  $\omega_{my}$  is month\*year fixed effects.

$hours\ worked_{i,mo \times y}$  are total hours worked by individual  $i$ , in month  $m$ , and year  $y$ ;  $X_i$  is a vector of controls, which includes age, age squared, gender, education, and an indicator if person is the household head. Results are presented in Tables CA.1 and CA.2.

**TABLE CA.1: The Many Uncertainties of Arsenic Contamination in Drinking Water**

**Health Outcomes**

Dependent variable: ihs (count of admissions)	1	2	3	4	5	6	(7	8
	Diarrhea	Abdominal pain	Vomiting	Dehydration	Hemolysis	Vertigo	Shock	Sum of all
Arsenic > 0.01	0.0481 (0.0387)	0.210** (0.103)	0.102** (0.0396)	0.0313* (0.0160)	-0.0484 (0.0501)	0.149 (0.108)	0.0653 (0.069)	0.301** (0.125)
Other controls	Total hospital admissions							
Month*year fixed effects	Y	Y	Y	Y	Y	Y	Y	Y
Province fixed effects	Y	Y	Y	Y	Y	Y	Y	Y
N	4067	4067	4067	4067	4067	4067	4067	4067
R <sup>2</sup>	0.134	0.617	0.384	0.152	0.254	0.370	0.325	0.687

Note: Standard errors in parentheses are clustered at the commune level. ihs = inverse hyperbolic sine. \*p < 0.10, \*\*p < 0.05, \*\*\*p < 0.01.

**Labor Outcomes**

	1	2	3
	Hours worked	Hours worked (Atacama region)	Pr(being employed) (Atacama region)
Arsenic > 0.01	-0.472*** -0.115	-1.974* -0.658	-0.056** (0.013)
Other controls	Age, age squared, gender, education, household head indicator		
Month*year fixed effects	Y	Y	Y
Municipality fixed effects	Y	Y	Y

Note: Robust standard errors are in parentheses. Pr = probability. \*p < 0.10, \*\*p < 0.05, \*\*\*p < 0.01.

## CHAPTER 3

### QUANTIFYING THE SENSITIVITY OF AGRICULTURAL PRODUCTION TO SALINITY

The estimating equation (C.13) for measuring the sensitivity of agricultural productivity to salinity is presented below.

Empirical strategy:

$$\Delta \log(NPP_{it}) = \alpha + \lambda * g(EC_{it}) + \delta * f(climate_{it}) + \sigma_i + \rho_y + \theta_c * Y + \varepsilon_{it} \quad (C.13)$$

$NPP_{it}$  is net primary productivity in grid cell  $i$  and year  $t$ ;  $EC_{it}$  is electrical conductivity monitoring station upstream from grid cell  $i$  and year  $t$ ;  $climate_{it}$  is quadratic terms of rainfall and temperature in grid cell  $i$  and year  $t$ ;  $\sigma_i$  is grid cell fixed effects;  $\rho_y$  is year fixed effects; and  $\theta_c * Y$  are state-specific time trends (or in the case of the India analysis, district-specific time trends). Identification is predicated on the upstream/downstream relationship of the monitoring station/grid cell that is described in box 1.1. Grid cells are included if they are within 100 kilometers from the upstream monitoring station and if the share of cropland in the grid cell is above 30 percent (columns 1 and 4), 75 percent (columns 2 and 5), and 90 percent (columns 3 and 6). Results are shown in Table C.11

**TABLE C.11: Quantifying the Sensitivity of Agricultural Production to Salinity**

	1	2	3	4	5	6	7	8	9	10
	Mekong River basin			India			GEMStat (global)			
EC > 100 mS/m	-0.058*** (0.0161)	-0.079*** (0.0187)	-0.082*** (0.0153)	-0.066*** (0.0123)	-0.056*** (0.00823)	-0.055*** (0.0108)	-0.112*** (0.0131)	-0.114*** (0.0197)	-0.135*** (0.0247)	
EC 40–80 mS/m										-0.074*** (0.00522)
EC 80–120 mS/m										-0.137*** (0.00703)
EC 120–160 mS/m										-0.257*** (0.0120)
EC 160–200 mS/m										-0.215*** (0.0128)
EC > 200 mS/m										-0.309*** (0.0266)
Other controls	Precipitation, precipitation squared, temperature, temperature squared									
Year fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Grid cell fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Geographic time trends	Country	Country	Country	State	State	State	Country	Country	Country	Country
Cropland threshold Percent	> 30	> 75	> 90	> 30	> 75	> 90	> 30	> 75	> 90	> 30
N	3,784	2,576	1,821	26,830	20,663	15,698	68,984	35,070	20,002	68,984
R <sup>2</sup>	0.674	0.645	0.62	0.476	0.478	0.467	0.237	0.275	0.314	0.244

Note: Standard errors in parentheses are calculated using two-way clustering, by monitoring station x year and province/state. EC = electrical conductivity.

\* $p < 0.10$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

## CHAPTER 4

### DETERMINANTS OF WATER QUALITY IN GLOBAL LAKES

The estimating equation (C.14) for measuring the determinants of water quality in lakes is presented below.

Empirical strategy:

$$WQ_{it} = \alpha + \alpha_1 \text{cropland}_{ut} + \alpha_2 \text{Log NPP}_{ut} + \alpha_3 \text{urban}_{ut} + \alpha_4 \text{Log pop}_{ut} + \alpha_5 \text{forest}_{ut} + \gamma \text{weather}_{(i+u)t} + \sigma_i + \rho_y + \text{country trends} + \epsilon_{it} \quad (\text{C.14})$$

$WQ_{it}$  is environmental water quality index for lake  $i$  and year  $t$ ;  $NPP_{ut}$  is net primary productivity and  $\text{pop}_{ut}$  is population in the upstream catchment area  $u$ ;  $\text{cropland}_{ut}$ ,  $\text{urban}_{ut}$ , and  $\text{forest}_{ut}$  are the shares of cropland, urban, and forest areas in the upstream catchment area  $u$ ;  $\text{weather}_{(i+u)t}$  are temperature and precipitation measures for the lake and upstream catchment region;  $\sigma_i$  are lake fixed effects; and  $\rho_y$  are year fixed effects.

Identification is predicated on the upstream/downstream relationship of the lake and catchment area where dependent variables are measured in upstream catchment areas and water quality is measured in downstream lakes. Results are shown in Table C.12.

**TABLE C.12: Determinants of Water Quality in Global Lakes.**

	Environmental water quality index	
Cropland (%)	3.5221168*	3.4194734*
	(2.021)	(2.013)
log NPP	0.0535863	
	(1.047)	
Urban (%)	3.4356898**	3.2916196**
	(1.660)	(1.655)
log population	3.7613908	
	(4.760)	
Forest (%)	0.5256785	0.4588324
	(0.714)	(0.680)
Average precipitation	0.0027260	0.0027657
	(0.002)	(0.002)
Precipitation squared	-0.0000006*	-0.0000006**
	(0.000)	(0.000)
Average temperature	0.4877382	0.4866587
	(0.383)	(0.384)
Temperature squared	-0.0015686	-0.0017362
	(0.014)	(0.014)
Lake fixed effects	Y	Y
Year fixed effects	Y	Y
N	2214	2223
R <sup>2</sup>	0.0732	0.0727

Note: Standard errors clustered at the lake level in parentheses.

\* $p < 0.10$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

## DETERMINANTS OF NITROGEN AND SALT IN GLOBAL RIVERS

The estimating equation (C.15) for measuring the determinants of water quality in lakes is presented below.

### Empirical strategy:

$$\begin{aligned}
 ihs(WQ_{it}) = & \alpha_1 + \alpha_2 * ihs(NPP_{it}) + \alpha_3 * urban_{it} + \alpha_4 * precip^+_{it} + \alpha_5 * precip^-_{it} + \alpha_6 * \\
 & temp_{it} + \alpha_7 * GDP_{it} + \alpha_8 * GDP^2_{it} + \sigma_i + \rho_y + \varepsilon_{it}
 \end{aligned}
 \tag{C.15}$$

$WQ_{it}$  is nitrates-nitrites (column 1) or electrical conductivity (column 2) in grid cell  $i$  and year  $t$ ;  $NPP_{it}$  is net primary productivity;  $urban_{it}$  is the share of urban area;  $precip^+$  and  $precip^-$  are indicators of precipitation one standard deviation above or below the long run mean for grid cell  $i$ ;  $temp_{it}$  is average annual temperature;  $GDP_{it}$  is local GDP;  $\sigma_i$  is grid cell fixed effects; and  $\rho_y$  is year fixed effects.  $ihs$  denotes the inverse hyperbolic

sine of the variable, an alternative to the natural log when the variable can take a value of 0.

Identification is predicated on the upstream/downstream relationship of the monitoring station/grid cell where dependent variables are measured in upstream grid cells and water quality is measured in downstream monitoring stations. Grid cells are included if they are within 100 kilometers from the downstream monitoring station. Results are shown in Table C.13.

**TABLE C.13: Determinants of Nitrogen and Salt in Global Rivers**

	1	2
	ihs (nitrates–nitrites)	ihs (electrical conductivity)
ihs (net primary productivity)	1.216*** (0.0327)	0.124*** (0.0107)
Share urban area	–0.955** (0.351)	0.541** (0.267)
ihs (population)	0.342*** (0.0477)	0.0247 (0.0247)
Positive precipitation shock	0.128*** (0.0148)	–0.0980*** (0.00783)
Negative precipitation shock	0.111*** (0.0114)	0.0387*** (0.0387)
Average annual temperature (C)	0.124*** (0.0106)	0.0118* (0.00550)
log (local GDP)	0.00664 (0.215)	0.797*** (0.0510)
log (local GDP) squared	0.0331*** (0.00660)	–0.0245*** (0.00173)
Grid cell fixed effects	Y	Y
Year fixed effects	Y	Y
N	21858	67507
R <sup>2</sup>	0.806	0.941

Note: Standard errors in parentheses. C = Celsius; GDP = gross domestic product; ihs = inverse hyperbolic sine.

\* $p < 0.10$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

## CHAPTER 6

### BOX C.3: INSTRUMENTS FOR WATER POLLUTION CONTROL— COMMAND AND CONTROL VERSUS ECONOMIC INCENTIVES

#### *Section 1: Corruption and Compliance under Alternative Policy Regimes*

Consider a tax and a pollution standard set such that *both* generate the *same* level of pollution under *perfect* compliance with the policies. Suppose that there is bribery: Will there be a difference in pollution (that is, compliance) levels generated by these two instruments?

To our knowledge, this issue has not been explored in the literature. We, therefore, develop a minimalist model as a metaphor to better understand the way in which different regulatory instruments induce corruption and noncompliance in regimes with imperfect enforcement. Although there is a wide and expanding range of environmental policies, such as technical mandates, voluntary restrictions, and tradeable rights in pollution, the focus of this paper is on taxes and standards, which remain the most widely used canonical regulatory instruments. We compare compliance incentives and pollution levels under taxes and standards and identify circumstances under which compliance can be expected to systematically differ between these instruments. The questions addressed in this paper are of critical relevance in the current context where environmental pressures are accelerating in developing countries and where enforcement capacity is limited.

The remainder of this chapter is as follows: Section 2 outlines the key results in a model with simple functional forms, and section 3 demonstrates that these results hold under general functional forms.

#### *Section 2: A Simple Model of Corruption and Environmental Regulation*

##### **The Benchmark Equilibrium**

For simplicity, consider a price-taking firm that produces an output ( $e$ ) that is sold in world markets at an exogenous price ( $p$ ). Let production costs be denoted by  $C(e)$ , with  $C'(e) > 0$  and  $C''(e) > 0$ . Assume that each unit of output generates an equivalent amount of pollution ( $e$ ), which causes environmental damage  $D(e)$  such that  $D'(e) > 0$  and  $D''(e) > 0$ . To generate closed-form solutions, we use simple functional forms in this section and specify the welfare function as

$$W = pe - \frac{ce^2}{2} - \frac{De^2}{2}. \quad (\text{C.16})$$

It is easily verified that in the absence of regulation, the profit-maximizing level of output (pollution) is defined by  $e^\pi = \frac{p}{c}$ , whereas the welfare-maximizing level of output (pollution) is given by  $e^w = \frac{p}{(c+D)}$ .

Suppose that the government seeks to regulate emissions through a linear tax on pollution.<sup>1</sup> The firm's profits are given by

$$\Pi^{ht} = pe - \frac{c}{2}e^2 - te. \quad (\text{C.17})$$

This implies that the profit-maximizing output level is  $e^\pi = \frac{p-t}{c}$ . To ensure that the firm produces at the welfare-maximizing level of emissions necessitates a tax rate of  $t^w = \frac{pD}{c+D}$ .<sup>2</sup>

Suppose instead that emissions are regulated through a pollution standard that is set at the welfare-maximizing level ( $e^s = e^w$ ). To ensure that profit-maximizing firms comply with the quota requires an enforcement mechanism that would create the incentives that are necessary to adhere to the standard. We consider a simple penalty regime and assume that when pollution levels exceed the regulatory standard, the firm incurs a fine (denoted  $V$ ) on the excess of emissions above the standard:  $V(e - e^w)$ . In this case, profits with a pollution standard are given by

$$\Pi^{hs} = pe - \frac{c}{2}e^2 - V(e - e^w) \quad (\text{C.18})$$

Observe that setting a penalty rate of  $V^w \geq t^w = \frac{pD}{c+D}$  will ensure that the firm remains in full compliance and never exceeds the welfare-maximizing standard ( $e^w$ ). Thus, in the absence of corruption, emissions taxes and a standard accompanied by a sufficiently severe fine for noncompliance yield equivalent outcomes.

### Imperfect Compliance and Corruption

We now explore whether the introduction of bribery alters this outcome. To compare equivalent cases, we use the welfare-maximizing outcome as the benchmark by comparing outcomes when pollution is regulated with a tax set at the welfare-maximizing level ( $t^w$ ) and when emissions are regulated through a standard set at the welfare-maximizing level ( $e^s = e^w$ ).

To ensure that results are not driven by arbitrary assumptions about differences in the penalties for noncompliance, the fine for exceeding the standard is set at a level of  $V = V^w = t^w$ , which imposes the same cost on the firm for noncompliance as it would have incurred with the welfare-maximizing tax ( $t^w$ )—in other words, when  $V^w = t^w$  exceeding the welfare-maximizing level of emissions imposes the same nominal compliance costs on the firm, regardless of whether the government chooses to regulate emissions with a tax or a standard.

We allow for the possibility of corruption by assuming that the regulatory agency does not directly observe emissions ( $e$ ) and, therefore, hires an inspector (a bureaucrat) who reports emission levels to the agency. When emissions are controlled through a pollution tax, the regulator imposes a tax ( $t = t^w$ ) on emissions that are *reported* by the inspector. We denote reported emissions under the tax by  $e_t^r$ .

The firm may seek to bribe the inspector an amount ( $B$ ) for reporting emissions ( $e_t^r$ ) that are lower than the true level of discharge ( $e$ ).<sup>3</sup> Thus, the actual tax burden on the firm under corruption is given by  $te_t^r$ , which may differ from that which would prevail in the absence of bribery ( $te$ ). Throughout this paper, we assume that  $e_t^r \leq e$ , which rules out the possibility of extortion—that the inspector reports higher emission levels than actually prevail.

Analogously, if a bribe is paid with a pollution quota, the inspector submits to the regulator a report ( $e_s^r \leq e^w$  — recall that  $e^w$  is the welfare-maximizing pollution standard) in exchange for a bribe ( $B$ ).

Although the regulator does not observe emissions, as in Laffont and Tirole (1998), it audits the firm based on a given prior about true pollution levels. Thus, with probability ( $\alpha$ ), an audit is initiated and uncovers true pollution levels ( $e$ ).<sup>4</sup> We follow judicial convention and assume that the fine (punishment) for underreporting emissions “fits the crime.” The firm and inspector are both fined an amount that is proportional to the level of underreporting of emissions. With a pollution tax, the fine on the firm is  $\frac{F}{2}(e - e_t^r)^2$ , and the inspector is fined  $\frac{I}{2}(e - e_t^r)^2$ . In the case of a pollution standard, the fines are respectively  $\frac{K}{2}(e - e^w)^2$  for  $K = F, I$  (where  $e^w$  is the welfare-maximizing standard as previously defined).

### Pollution Tax

Consider first the case of a pollution tax and suppose that the firm decides to offer the inspector a bribe ( $B > 0$ ) to report a lower level of emissions ( $e_t^r < e$ ). Let  $\Pi^h$  denote firm profits from honest revelation of emissions when there is no payment of bribes. The expected gains from offering a bribe are.

$$U_t^F = pe - \frac{c}{2}e^2 - te_t^r - \frac{\alpha F}{2}(e - e_t^r)^2 - B - \Pi^h. \quad (\text{C.19})$$

In equation (C.19), the first term ( $pe$ ) represents the total revenue from production, the second term ( $\frac{c}{2}e^2$ ) the costs of production, and the third term ( $te_t^r$ ) the tax on reported emissions. Submitting a lower report ( $e_t^r < e$ ) brings additional noncompliance costs: With probability ( $\alpha$ ), the firm is audited and fined ( $F$ ) in accordance with the magnitude of underreporting ( $e - e_t^r$ ). In addition, a bribe ( $B$ ) must be paid to induce the inspector to underreport true emissions ( $e_t^r < e$ ).

Likewise, the expected gains to the inspector are

$$U_t^I = w + B - \frac{\alpha I}{2}(e - e_t^r)^2 - w, \quad (\text{C.20})$$

where  $w$  is the wage paid to the inspector,  $B$  is the bribe, and the third term is the expected fine. The final term ( $w$ ) is the payoff from not accepting a bribe and being honest.

As is usual in the corruption literature, reported and actual emissions are determined to maximize the joint payoffs of both parties:

$$\text{Max}_{e, e_t^r} U_t = U_t^F + U_t^I \quad (\text{C.21})$$

The first order conditions for  $e$  and  $e_t^r$  are, respectively,

$$p - ce - \alpha\Omega(e - e_t^r) = 0, \tag{C.22}$$

$$-t + \alpha\Omega(e - e_t^r) = 0, \tag{C.23}$$

where  $\Omega = (F + I)$  is the expected fine on both parties. Solving, we obtain reported and actual emission levels:

$$e^t = \frac{p-t}{c} \text{ and } e^r = \frac{p-t}{c} \tag{C.24}$$

Figure C.3.1 provides a graphical summary of the outcome. From equation (C.24), it is clear that reported emissions rise linearly with actual emissions, though the inspector underreports true emission levels by an amount of  $\frac{t}{\alpha\Omega}$ . The reported emissions function is shown in figure C.3.1, panel a. Figure C.3.1, panel b illustrates the equilibrium level of output, which is determined by the intersection of the (gross) marginal payoffs from production ( $p - ce$ ) and the expected marginal costs of non-compliance ( $\alpha\Omega(e - e_t^r)$ ) in equation (C.22). Moreover, by equation (C.23) in equilibrium, the firm and inspector optimally trade off the benefit of underreporting emissions, which lowers the tax burden against the costs of a fine if audited.

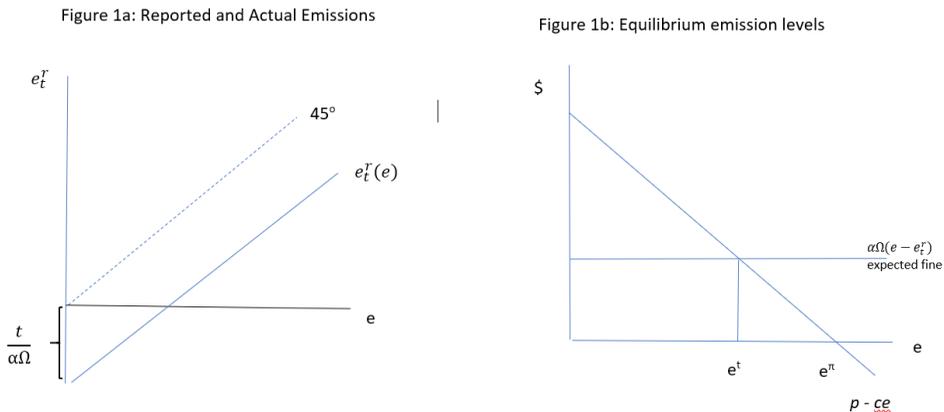
### Pollution Standard

Consider next the case where the regulator imposes a pollution standard at the welfare-maximizing level ( $e^w$ ). The firm's payoffs from offering a bribe to report a lower level of emissions ( $e_s^r$ ) are given by

$$U_s^F = pe - \frac{c}{2}e^2 - \frac{\alpha F}{2}(e - e_s^r)^2 - \delta V(e_s^r - e^w) - B - \Pi^{hs}, \tag{C.24}$$

where  $\delta$  is an indicator variable that is equal to unity when  $e_s^r - e^w > 0$  and is 0 otherwise.

**FIGURE C.3.1:** Graphical Summary



Equation (C.24) is analogous to equation (C.22). The first two terms describe the revenue and production costs, respectively, from an output level of  $e$ . The third term represents the fine payable by the firm when successfully audited for an untrue report ( $e_s^r < e$ ). The fourth term defines the penalty incurred on reported emissions in excess of the welfare-maximizing standard ( $e^w$ ).  $B$  is the bribe paid to the inspector, and  $\Pi^{hs}$  are payoffs in the absence of corruption.

Similarly, the inspector's payoffs are described by

$$U_s^I = w + B - \frac{\alpha^I}{2}(e - e_s^r)^2 - w. \quad (\text{C.25})$$

Combining equations (C.24) and (C.25), the joint payoffs are given by

$$U_s = pe - \frac{c}{2}e^2 - \frac{\alpha\Omega}{2}(e - e_s^r)^2 - \delta V(e_s^r - e^w), \quad (\text{C.26})$$

where  $\Omega = (F + I)$ .

Formally, this problem requires that the inspector and firm select the output ( $e$ ) and report ( $e_s^r$ ) to maximize joint payoffs ( $U_s$ ), subject to the constraint that  $e_s^r \geq e^w$ . Because the constraint may not always bind, from the Kuhn-Tucker conditions, this implies that a number of alternative cases need to be considered.

Equation (C.26) suggests that three equilibria are possible. First, the solution may yield an outcome where reported emissions exceed the standard ( $e^w$ ):  $e_s^r > e^w$ . In this case, the firm incurs a fine of  $V$  on reported emissions in excess of the standard. A second solution entails a report ( $e_s^r = e^w$ ). In this case, the inspector reports that the firm is in compliance, so it avoids the fine on reported emissions. Thus, in equation (C.26),  $\delta = 0$ .<sup>5</sup> A final outcome that we rule out is that the inspector submits a report ( $e_s^r < e^w$ ). This can be shown to never be optimal. Intuitively, submitting a report below the standard ( $e^w$  (i.e.,  $e_s^r < e^w$ )) is never optimal because it confers no benefit through a false claim that emissions are below the standard, but it unambiguously increases the expected penalty on both the firm and the inspector.<sup>6</sup>

Consider first the case when the constraint does not bind and  $e_s^r > e^w$ . The first-order conditions for  $e$  and  $e_s^r$  are

$$p - ce - \alpha\Omega(e - e_s^r) = 0, \quad (\text{C.27})$$

$$\alpha\Omega(e - e_s^r) - V = 0. \quad (\text{C.28})$$

Solving, we obtain

$$e_{s1}^r = e - \frac{V}{\alpha\Omega} \text{ and } e_{s1} = \frac{p-V}{c}. \quad (\text{C.29})$$

Consider next the equilibrium with the binding constraint. The first order conditions are

$$p - ce - \alpha\Omega(e - e^w) = 0, \quad (\text{C.30})$$

$$(e^w - e_s^r) = 0, \quad (\text{C.31})$$

with output and reported emissions given by

$$e_{s2}^r = e^w, \text{ and } e_{s2} = \frac{p + \alpha \Omega e^w}{c + \alpha \Omega} \quad (\text{C.32})$$

Lemmas C.1 and C.2 compare these equilibria with those that obtain with a pollution tax. Consider first the case when the constraint binds. Comparing equations (C.32) and (C.24) shows that when it is optimal for the inspector and firm to report full compliance with the standard, emission levels under the standard will always exceed those under the tax.

LEMMA C.1. *Comparing equations (C.32) and (C.24), whenever  $e_{s2}^r = e^w$ , then  $e_{s2} - e_t = \frac{t^w c}{c(c + \Omega)} > 0$ .*<sup>7</sup>

In sum, lemma C.1 indicates that when it is optimal to report that the firm is in compliance with the emissions standard, then pollution levels under the standard exceed those with a pollution tax. Thus, a pollution tax induces greater compliance levels (as defined by the level of pollution) than does a standard.

LEMMA C.2. *Comparing equations (C.29) and (C.24), whenever  $V = t^w$  and  $e_{s1}^r > e^w$  then  $e_{s1} = e_t$ .*

Lemma C.2 compares outcomes when the constraint does not bind and when  $V = t^w$ —a situation where the fine for reporting emissions in excess of the standard imposes the same cost as a pollution tax at the optimum rate  $t^w$ . Substituting  $V = t^w$  in equation (C.29) makes clear that in this outcome, both the pollution standard and the pollution tax yield an identical outcome. Intuitively, this outcome is a consequence of the overall penalties for noncompliance being the same under both instruments when noncompliance is reported.

However, it is unclear whether, and in what contexts, the inspector will or will not report full compliance with the standard (that is, whether the constraint does or does not bind). Lemma C.3 explores this issue in more detail and asks which of these outcomes is most likely to prevail. It finds that in regimes with low-enforcement capabilities (as defined by a small  $\alpha$ —the probability of a successful audit and prosecution), it always pays to report full compliance. The reason is perhaps obvious. In low-compliance contexts, the expected penalty for underreporting true emissions will be low, so it always pays to report full compliance with a standard and risk a fine (which will occur with low enough probability).

LEMMA C.3. *At any given level of emissions,*

*$e \in (e^w, e^\pi]$   $U_s(e_s^r = e^w) - U_s(e_s^r > e^w) > 0$  whenever*

*$\alpha < \frac{2V}{\Omega(2e - e_s^r - e^w)} \equiv \alpha^*$ . That is, the joint payoffs from reporting full*

*compliance ( $U_s(e_s^r = e^w)$ ) exceed the joint payoffs from reporting*

*noncompliance ( $U_s(e_s^r > e^w)$ ) whenever  $\alpha$  is sufficiently small.*

PROOF. Define  $U_1(e_s^r > e^w) = pe - \frac{c}{2}e^2 - \frac{\alpha\Omega}{2}(e - e_s^r)^2 - \delta V(e_s^r - e^w)$   
 and  $U_2(e_s^r = e^w) = pe - \frac{c}{2}e^2 - \frac{\alpha\Omega}{2}(e - e^w)^2$ . For any given  $e \in (e^w, e^\pi]$ ,  
 $U_1(e_s^r > e^w) - U_2(e_s^r = e^w) = (e_s^r - e^w)(e\alpha\Omega - \frac{\alpha\Omega(e_s^r + e^w)}{2})$   
 $-V) < 0$  if  $\alpha < \frac{2V}{\Omega(2e - e_s^r - e^w)} \equiv \alpha^*$ .

Not implausibly, lemma C.3 indicates that in low-enforcement contexts that typify developing countries, there is little incentive to report noncompliance when it occurs. In such circumstances, lemma C.1 indicates that compliance under a standard will be lower than under a tax. Intuitively, with a standard, the only penalty the firm faces is a fine for underreporting, which is incurred with low probability, whereas with a tax, the firm incurs both the low-probability penalty for underreporting as well as a levy on all (inframarginal) reported levels of pollution. Thus, for any given level of underreporting, the overall regulatory burden with a tax (comprising the tax on reported emissions and expected fine for underreporting) is higher than with a standard (which comprises only the expected fine for underreporting).

### ***Section 3: Generalization of the Results***

This section explores whether the results are a consequence of the specific functional forms used and whether they hold under more general conditions. The results suggest that under seemingly mild assumptions that generate interior solutions, the results hold in contexts of weak regulation that are conducive to corruption. Because enforcement capacity is weakest in developing countries, where much global and local environmental damage is concentrated and accelerating, the implication of this result is that the environmental policy production function in developing country contexts may differ in significant ways from those in high-income economies with the necessary enforcement capabilities.

For consistency with the model presented in the previous section, welfare is defined as

$$W(e) = pe - C(e) - D(e). \quad (C.33)$$

Without loss of generality, we retain the assumption that a unit of output ( $e$ ) generates an equivalent amount of pollution. In addition, price is assumed exogenous because the good is traded on world markets; production costs are denoted  $C(e)$ , with  $C'(e) > 0$  and  $C''(e) > 0$ ; and environmental damage is  $D(e)$  such that  $D'(e) > 0$  and  $D''(e) > 0$ .

Analogously, under taxation, the joint payoff function maximized by the firm and inspector is

$$\text{Max}_{e, e_t^r} U_t = pe - C(e) - te_t^r - \alpha\Omega(e - e_t^r), \quad (C.34)$$

where  $\Omega(e - e_t^r)$  is the joint penalty for underreporting, which is assumed to be nondecreasing in the level of underreporting. Thus,  $\Omega'(e - e_t^r) \geq 0$  and  $\Omega''(e - e_t^r) \geq 0$ . The first-order conditions for  $e$  and  $e_t^r$  are, respectively,

$$p - C' - \alpha\Omega'(e - e_t^r) = 0, \quad (\text{C.35})$$

$$\alpha\Omega'(e - e_t^r) - t = 0. \quad (\text{C.36})$$

In a similar way, the maximization problem with a pollution standard is defined by

$$\text{Max}_{e, e_s^r} U_s = pe - C(e) - \delta V(e_s^r - e^w) - \alpha\Omega(e - e_t^r) \text{ subject to } e_s^r \geq e^w, \quad (\text{C.37})$$

whereas in the previous section,  $V'(e_s^r - e^w) > 0$  and  $V''(e_s^r - e^w) = 0$ . Focusing upon the interior solutions (where  $e > 0$  and  $e_s^r \geq 0$ ), the Kuhn-Tucker necessary conditions are given by

$$p - C'(e) - \alpha\Omega'(e - e_s^r) = 0, \quad (\text{C.38})$$

$$\alpha\Omega'(e - e_s^r) - V'(e - e_s^r) - \lambda = 0, \quad (\text{C.39})$$

$$\lambda(e_s^r - e^w) = 0 \rightarrow \lambda = 0, \text{ if } (e_s^r - e^w) > 0 \text{ and } \lambda \geq 0, \text{ if } (e_s^r - e^w) = 0. \quad (\text{C.40})$$

The following lemmas describe the main results with all proofs relegated to annex CB.

As in the previous section, lemma C.4 compares the equilibrium with a standard and a tax when the constraint binds  $e_s^r = e^w$  and finds that in contexts with a low probability of a successful audit, emission levels are higher under a standard. Although this result was found to always hold with the specific functional forms used in the previous section, in general, it can be expected to occur only when enforcement capacity is weak (low  $\alpha$ ). In such circumstances, the expected cost of noncompliance under a standard is lower than with a pollution tax.

**LEMMA C.4.** *For sufficiently small  $\alpha$ , and when  $e_s^r = e^w$ , then  $e_s > e_t$ . That is, when  $e_s^r = e^w$  and  $\alpha < \frac{t}{\Omega'(e_s - e^w)} \equiv \bar{\alpha}$ , then  $e_s > e_t$ .*

Lemma C.5 demonstrates that in low-enforcement regimes, it never pays to report noncompliance.

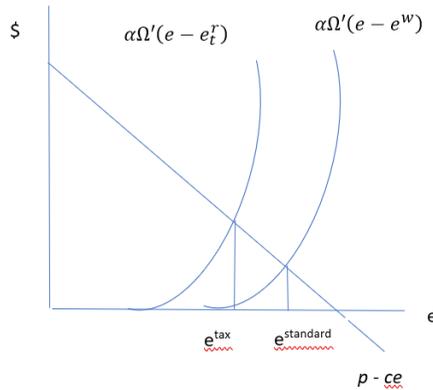
**LEMMA C.5.** *For sufficiently small  $\alpha$ , it always pays to report compliance with the standard ( $e_s^r = e^w$ ). That is, for a given  $e_s \in (e^w, e^\pi]$ ,  $U_s(e_s^r = e^w) - U_s(e_s^r > e^w) > 0$  if  $\alpha < \frac{V(e_s^r - e^w)}{-\Omega(e_s - e_s^r) + \Omega(e_s - e^w)} \equiv \bar{\alpha} > 0$ .*

Taken together, lemmas C.4 and C.5 imply the following result:

**Result.** When  $\alpha \leq \min(\bar{\alpha}, \bar{\alpha})$ , the inspector submits a report of full compliance and emissions under the standard exceed those under a pollution tax. That is, when  $\alpha \leq \min(\bar{\alpha}, \bar{\alpha})$ , then  $e_s^r = e^w$  and  $e_s > e_t$ .

This result is perhaps intuitive. When enforcement capacity is sufficiently low, there is little incentive for an inspector to report noncompliance with a standard (lemma C.5). With a tax, even when emissions are underreported, the firm must pay tax on the reported emissions. Hence, the expected marginal payoffs from production and pollution are higher with a standard than with a tax (lemma C.4). As a result, in low-enforcement regimes that typify developing country situations, an environmental standard induces greater corruption and emissions than does a tax. The resulting low-compliance (small  $\alpha$ ) equilibria are illustrated in figure C.3.2. The policy implication of this finding is clear, and it calls for a greater focus of empirical and theoretical research on the rent-seeking opportunities and compliance incentives generated by different environmental instruments.

**FIGURE C.3.2:** Comparison of Low-Compliance Equilibria



## ANNEX CB

### *Proof of Lemma C.4*

Let  $e_t$  be the equilibrium pollution (output) levels with a tax and  $e_s$  the corresponding pollution levels with standard. Suppose that  $e_s > e_t$ ; From the first-order conditions and concavity of payoffs, this implies that  $p - C'(e_t) < p - C'(e_s) \rightarrow \alpha\Omega'(e_s - e^w) < t$ , which holds whenever  $\alpha < \frac{t}{\Omega'(e_s - e^w)} \equiv \bar{\alpha}$ . Alternatively, observe that as  $\alpha \rightarrow 0$ , then  $\alpha\Omega'(e_s - e^w) \rightarrow 0$ .

### *Proof of Lemma C.5*

Consider some  $e_s \in (e^w, e^\pi]$ . Note that

$$U_s(e_s^r = e^w) - U_s(e_s^r > e^w) = V(e_s^r - e^w) - \Omega(e_s - e_s^r) + \Omega(e_s - e^w)$$

. Upon rearrangement,  $U_s(e_s^r = e^w) - U_s(e_s^r > e^w) > 0$  if

$$\alpha < \frac{V(e_s^r - e^w)}{-\Omega(e_s - e_s^r) + \Omega(e_s - e^w)} \equiv \bar{\alpha} > 0 .$$

## REFERENCES

Laffont, J. J., and J. Tirole. 1998. *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press.

## BOX C.4: TREATING WASTEWATER: END-OF-PIPE SOLUTIONS OR PIPE DREAMS?

Two main estimation strategies are utilized to test for the effectiveness of wastewater treatment plant (WWTP) construction. First, a simple difference-in-difference model is estimated, where any monitoring station within a set distance downstream of a WWTP is considered treated, while all other monitoring stations are considered controls. Second, an event-study analysis is conducted for each WWTP. In this specification, monitoring stations downstream of a WWTP are compared to monitoring stations upstream both before and after each WWTP is constructed. In all of the following specifications, the estimated effect of wastewater treatment should be interpreted as the average across all treatment plants. Considerable heterogeneity in this treatment likely exists due to unobserved factors, such as river discharge rate, local pollution sources not captured by the wastewater treatment infrastructure, and pollution inflow from nearby tributaries, none of which are observed in the data.

The difference-in-difference approach employed here is similar to the methodology used by Greenstone and Hanna (2014). The estimated model is:

$$y_{it} = \sum_s (\beta_1 T_{ist} + \beta_2 T_{ist}t + \beta_3 D_{ist}) \mathbf{1}[d(i,s) < \bar{d}] + \alpha_i + \gamma_t + \varepsilon_{it} \quad (C.41)$$

The outcome variable  $y_{it}$  is the log pollution level measured at monitoring station  $i$  in year  $t$ . The treatment variable  $T_{ist}$  is equal to one if monitoring station  $i$  is downstream of STP  $s$  in year  $t$ , whereas the indicator  $D_{ist}$  is equal to one if STP  $s$  is ever observed upstream of  $i$  in the sample.

The indicator function  $\mathbf{1}[\cdot]$  filters the treatment variable by downstream distance. Since the predicted effect of an STP on downstream pollution levels diminishes as distance from the STP increases, it is necessary to choose a downstream distance  $\bar{d}$  beyond which the STP is predicted to have no effect. In the data, the function  $d(i,s)$  is the great-circle distance between monitoring station  $i$  and upstream STP  $s$ . The selection of  $\bar{d}$  must be weighed against the geospatial sparseness of the pollution data—small choices for  $\bar{d}$  will result in few treated monitoring stations, while large  $\bar{d}$  will introduce heterogeneity in the predicted treatment effect for treated monitoring stations nearby and STP compared to those further downstream. For robustness, various bandwidths ( $\bar{d} \in \{10km, 25km, 50km\}$ ) are chosen for each specification.

Monitoring station specific heterogeneity is captured by the fixed effect  $\alpha_i$ . This term captures all factors local to each monitoring station that are not time varying. The term  $\gamma_t$  is a year fixed effect which captures time varying factors that affect all monitoring stations equally, such as weather patterns or macroeconomic conditions.

The primary parameters of interest are  $\beta_1$  and  $\beta_2$ . As the coefficient on the treatment variable  $T_{ist}$ , the parameter  $\beta_1$  captures (the log approximation of) the percentage change in pollution levels that occurs downstream of an STP following its construction.  $\beta_2$  accounts for changes in pollution trends that are caused by STPs. Negative estimates of  $\beta_1$  and  $\beta_2$  would be consistent with a reduction in pollution levels caused by STP construction.

The term  $\beta_3 D_{ist}$  allows for monitoring stations that are downstream of an STP to have different trends than upstream, untreated monitoring stations. Therefore, the estimates of  $\beta_3$  serve as an explicit test of the “parallel trends assumption” that is common in the difference-in-difference framework.

Standard errors are clustered at the level of the monitoring station and year. This allows for errors to be within each monitoring station and across monitoring stations within each year. The correlation of errors in the true data generating process are more complex than can be captured by a postestimation clustering strategy, however. Errors are likely to be correlated among stations nearby one another along the same river system, but this correlation likely shrinks as downstream distance between stations increases. The choice of estimating monitoring station and year clusters was motivated primarily by a desire to yield conservative estimates of statistical significance. In practice, there appears to be little numerical difference between standard errors clustered at various levels (watershed or state, for instance). Additionally, the statistical inference based on the clustered standard errors is broadly similar to the inference conducted with simple heteroskedastic-robust standard errors.

Results for China and India are reported in Table C.4.1 and Table C.4.2. Equation (C.41) is estimated at three different bandwidths. In the Chinese data, there are no monitoring stations within 10km downstream of any WWTP, so the effects at this bandwidth cannot be estimated. For each pollutant, the first column in the table (e.g., Columns 1, 4, and 7) report the results for a simple difference-in-difference estimation with no time trends ( $\beta_2 = \beta_3 = 0$ ), while the second column adds aggregate time trends. The last column adds individual time trends for each monitoring station which allows pollution trends to vary locally.

For India, an event study analysis is also conducted using an approach utilized by Keiser and Shapiro (2018) for their study in the United States. Pollution levels are compared for all monitoring stations within a radius  $r$  (in kilometers) upstream and downstream of a particular STP, with upstream monitoring stations serving as the control and downstream as treated. As  $r$  shrinks to zero, any treatment externalities caused by upstream STPs will affect all monitoring stations within  $r$  equally. In a sense, this approach shifts the unit of observation from the individual monitoring stations to the local region around each STP.

**TABLE C.4.1: Difference-in-Difference Estimation Results, China**

	Dependent Variable					
	log(nh3)			log(cod)		
	(1)	(2)	(3)	(4)	(5)	(6)
<b>Panel A: <math>d(i, s) &lt; 25\text{km}</math></b>						
$T_{ist}$	-0.009 (0.009)	-0.005* (0.003)	-0.0001 (0.013)	-0.009* (0.005)	-0.010 (0.007)	-0.012 (0.011)
$D_{is} \cdot t$		0.004 (0.006)			-0.004 (0.009)	
$T_{ist} \cdot t$		-0.004 (0.006)	-0.009 (0.006)		0.004 (0.009)	0.005 (0.007)
<b>Panel C: <math>d(i, s) &lt; 50\text{km}</math></b>						
$T_{ist}$	-0.006 (0.009)	-0.024** (0.012)	-0.017 (0.013)	0.007* (0.004)	0.006*** (0.002)	-0.001 (0.007)
$D_{is} \cdot t$		-0.012 (0.009)			-0.001 (0.003)	
$T_{ist} \cdot t$		0.012 (0.009)	0.004 (0.008)		0.001 (0.003)	0.001 (0.004)
Year Fixed Effects	Yes	Yes	No	Yes	Yes	No
Station Specific Trends	No	No	Yes	No	No	Yes
Observations	1,332	1,332	1,332	1,292	1,292	1,292
$R^2$	0.822	0.822	0.902	0.879	0.879	0.938

Note: All regressions include monitoring station fixed effects. Monitoring station by year standard errors in parentheses.

\* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

**TABLE C.4.2: Difference-in-Difference Estimation Results, India**

Dependent Variable									
	log(bod)			log(cod)			log(fcoli)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
<b>Panel A: d(i, s) &lt; 10km</b>									
<i>Tist</i>	0.029	-0.083	-0.024	0.103**	0.062	-0.013	0.374***	-0.050	0.094
	(0.036)	(0.052)	(0.062)	(0.042)	(0.076)	(0.059)	(0.101)	(0.178)	(0.237)
<i>Dis · t</i>		0.006***			0.003			0.029***	
		(0.002)			(0.002)			(0.008)	
<i>Tist · t</i>		0.0004	-0.001		-0.0002	0.002		0.0003	0.007
		(0.002)	(0.002)		(0.003)	(0.002)		(0.007)	(0.009)
<b>Panel B: d(i, s) &lt; 25km</b>									
<i>Tist</i>	0.015	-0.099***	-0.056	0.065**	0.039	-0.005	0.275***	-0.233**	0.093
	(0.02)	(0.037)	(0.049)	(0.029)	(0.054)	(0.044)	(0.077)	(0.111)	(0.199)
<i>Dis · t</i>		0.006***			0.002*			0.018***	
		(0.001)			(0.001)			(0.006)	
<i>Tist · t</i>		-0.0002	-0.001		-0.001	0.002		0.008	0.009
		(0.001)	(0.002)		(0.003)	(0.002)		(0.005)	(0.008)
<b>Panel C: d(i, s) &lt; 50km</b>									
<i>Tist</i>	0.015	-0.039	-0.010	0.052**	0.044	0.005	0.257***	-0.073	0.022
	(0.013)	(0.027)	(0.031)	(0.020)	(0.037)	(0.036)	(0.051)	(0.078)	(0.125)
<i>Dis · t</i>		0.005***			0.003**			0.011***	
		(0.001)			(0.001)			(0.004)	
<i>Tist · t</i>		-0.002	-0.001		-0.002	-0.0002		0.005	0.011**
		(0.001)	(0.001)		(0.002)	(0.003)		(0.004)	(0.006)
Year Fixed Effects	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No
Station Specific Trends	No	No	Yes	No	No	Yes	No	No	Yes
Observations	19,812	19,812	19,812	8,765	8,765	8,765	14,310	14,310	14,310
R <sup>2</sup>	0.796	0.797	0.854	0.740	0.740	0.806	0.748	0.749	0.840

Notes: All regressions include monitoring station fixed effects. Monitoring station by year standard errors in parentheses.

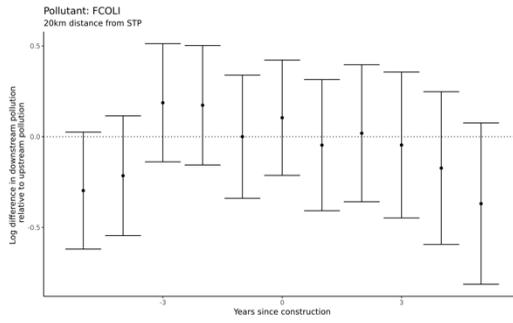
Formally, the following equation is estimated:

$$y_{ist} = \beta_1 T_{ist} + \beta_2 \tau + \beta_3 D_{ist} + T_{ist} \tau + \alpha_i \mu_s + \gamma_\tau \mu_s + \varepsilon_{ist} \quad (\text{C.42})$$

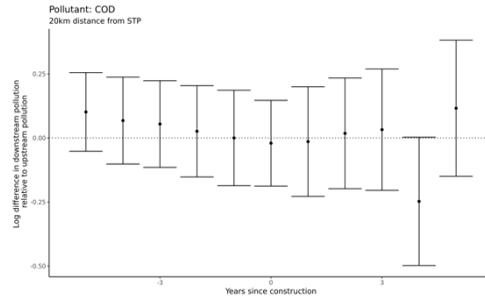
Here the outcome variable  $y_{ist}$  is the pollution level at monitoring station  $i$  nearby (within  $r$  kilometers) of STP  $s$  (20 kilometers in the results below). For each STP, pollution observations  $\tau$  years after the STP is constructed are recorded ( $-5 \leq \tau \leq 5$ ), with  $\tau = 0$  corresponding to the year in which the STP is constructed. This specification compares monitoring stations just downstream of an STP with those just upstream of the same STP. The term  $\gamma_\tau \mu_s$  captures year-to-year factors that affect pollution levels at all the monitoring stations local to a particular STP. This accounts for the effects local weather patterns, economic conditions, or regulatory structure that varies across locations within a country.

Figures C.4.1 through C.4.3 show the change in relative downstream pollution as a percentage of the change in upstream pollution before and after WWTP construction. None of the three pollutants and water quality indicators show any meaningful decrease after an WWTP is constructed relative to upstream pollution levels.

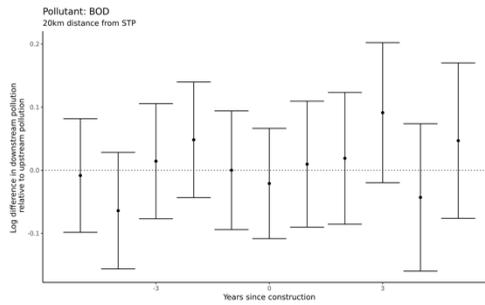
**FIGURE C.4.1: Fecal Coliform**



**FIGURE C.4.2: COD**



**FIGURE C.4.3: BOD**



## NOTES

1. A linear tax is arguably more realistic, whereas if a quadratic tax is considered, at the optimum, it is simply set equal to  $D$ .
2. To see why, substitute  $tw$  into  $e^t$  to obtain  $ew$ .
3. We model how reported emissions are determined in what follows.
4. We exclude learning and signaling as determinants of the audit probability because these issues are parenthetical to the concerns of this paper and complicate the analysis without adding relevant insights.
5. There is, of course, a tradeoff between submitting a low report that avoids the fine ( $V$ ) for exceeding the standard and the higher penalty that would be incurred with a lower report if there were a successful audit.
6. More formally, to see why, let  $e1 < ew$ . For any given output level ( $e$ ), if the firm and inspector submit  $e_1$  as the reported level of emissions, they incur an expected penalty of  $R_1 = \frac{\alpha\Omega}{2}(e - e_1)^2$ ; if they were to submit a report ( $ew > e1$ ), the expected penalty is  $R_2 = \frac{\alpha\Omega}{2}(e - e^w)^2$ . Clearly, because  $e1 < ew$ , then the expected penalty  $R_1 > R_2$ . Hence, it is never optimal to set  $e_s^r < e^w$ .
7. In fact, there is an additional term:  

$$e_{s2} - e_t = \frac{t^w c + \alpha\Omega(e^w c - p + t^w)}{c(c+\Omega)} > 0$$
because at  $tw$ , we have from from the first order conditions of equations (C.23) and (C.24)  

$$e^w c - p + t^w = 0$$

## REFERENCES

- Keiser, D. A., and J. S. Shapiro. 2018. "Consequences of the Clean Water Act and the Demand for Water Quality." *Quarterly Journal of Economics* 134 (1): 349–96.
- Greenstone, M., and R. Hanna. 2014. "Environmental Regulations, Air and Water Pollution, and Infant Mortality in India." *American Economic Review* 104 (10): 3038–72. doi:10.1257/aer.104.10.3038.