The Costs of Industrial Water Pollution to Agriculture in India

Nick Hagerty[*]

May 13, 2018

Abstract

Industrial water pollution is high in many developing countries, but researchers and regulators have paid it less attention than air and domestic water pollution. I estimate the costs of industrial water pollution to agriculture in India, focusing on 63 industrial sites identified by the central government as "severely polluted." I exploit the spatial discontinuity in pollution concentrations that these sites generate along a river. First, I show that these sites do in fact coincide with a large, discontinuous rise in pollutant concentrations in the nearest river. Then, I estimate that agricultural revenues are nine percent lower in districts immediately downstream of polluting sites, relative to districts immediately upstream of the same site in the same year, although confidence intervals exclude zero only when controlling for baseline characteristics. This effect appears to be driven by reduced yields per cropped land area and not factor reallocation. These results suggest that damages to agriculture could represent a major cost of water pollution and warrant further study.

[*] Department of Economics, Massachusetts Institute of Technology. I thank Ben Olken and Esther Duflo for advice, as well as Simon Jäger, Abhijit Banerjee, Anant Sudarshan, Peter Hull, and Alex Oberg for helpful comments and discussions. I am grateful for financial support from the MIT Tata Center for Technology & Design.

1 Introduction

Pollution levels in developing countries are often orders of magnitude larger than in developed countries. Simple linear extrapolation suggests that the costs to health, productivity, compensatory behavior, and ecology are relatively more important in developing countries. Moreover, there is evidence that costs from pollution may be nonlinear, with marginal costs increasing in pollution levels (Arceo et al. 2016). Unfortunately, most quasi-experimental evidence on the costs of pollution comes from developed countries, with little basis to extrapolate to developing settings. More specifically, air pollution has enjoyed growing academic interest and exploding public interest in developing countries, but water pollution, while also high, has received less attention. Public pressure in India has brought down air pollution levels over the last few decades (to still-extreme levels by developed country standards), but similarly strict regulation has not discernibly improved water quality (Greenstone and Hanna 2014).

In this paper I estimate the costs of industrial water pollution to agriculture in India, taking advantage of an unusally rich developing country dataset on water pollution obtained and made public by Greenstone and Hanna (2014). I focus on 63 industrial sites identified by the Central Pollution Control Board in 2009 as “severely polluted” with respect to water pollution, out of 88 sites selected for intensive study. Unlike other forms of pollution, water pollution almost always flows in only one direction from its source. My empirical approach exploits the fact that when industrial wastewater is released into a flowing river, it creates a spatial discontinuity in pollution concentrations along that river. Areas immediately downstream of a heavily polluting industrial site will have higher pollution levels than areas immediately upstream, yet they are likely similar otherwise. This makes upstream areas a reasonable counterfactual for the downstream areas in studying the impacts of water pollution on economic outcomes.

I show three sets of results. First, I show that there is a large, discontinuous increase in industrial pollution at the location of these “severely polluted” industrial sites which raises existing levels of pollution in the rivers by a factor of three to five. Second, I find that agricultural revenues in districts immediately downstream of these industrial sites are 9 percent lower than in upstream districts, with a 95% confidence interval of 2 to 17 percent. Third, I document that farmers are neither substituting factors of production away from agriculture nor applying additional compensating inputs; the effect of being downstream on yields is large and significant, while the effects on crop area, irrigation, fertilizer, labor, and population are small to zero. Two important caveats are that these effects are precise enough to exclude zero only when accounting for baseline characteristics, and that this approach cannot exclude the possibility that areas upstream and downstream of pollution sources differ in other systematic ways. Still, these results suggest that damages to agriculture may represent a major cost of water pollution and warrant further study.

This paper contributes to the existing literature in four ways. First, it provides among the first quasi-experimental evidence on any form of costs of industrial water pollution. In recent papers on India, Do et al. (2018) study the effects of industrial water pollution on infant mortality, while Brainerd and Menon (2014) study the effects of agricultural water pollution to child health. In the United States, Keiser and Shapiro (2017) study the effect of all water pollution on property values. Most other economics literature on the costs of water pollution deals with domestic water pollution in the context of providing clean drinking water. Second, this paper documents economic costs of pollution to agriculture; to my knowledge only one other paper does so quasi-experimentally, but in the context of air pollution (Aragón and Rud 2016). Third, it adds to the small but rapidly growing literature on the costs of pollution in developing countries (Jayachandran 2009; Chen et al. 2013; Adhvaryu et al. 2016).

Finally, this paper contributes to a broader understanding of structural transformation and the relationship between industry and agriculture in developing countries. Most existing literature focuses on input reallocation between sectors (Ghatak and Mookherjee 2014; Bustos et al. 2016), while this paper is among the first to document a non-pecuniary externality from industry to agriculture.

2 Background

2.1 Industrial water pollution and crops

Manufacturing plants like those in India produce a variety of waste chemicals which, if untreated or insufficiently treated, will reach surface or ground water systems. These chemicals include organic chemicals including petroleum products and chlorinated hydrocarbons; heavy metals including cadmium, lead, copper, mercury, selenium, and chromium; salts and other inorganic compounds and ions; and acidity or alkalinity. Many of these products are carcinogenic or otherwise toxic in sufficient quantities to humans and other plants and animals.

Agricultural crops are no exception. Biologically, it is well known that plant growth is sensitive to salinity, pH (i.e., acidity and alkalinity), heavy metals, and toxic organic compounds. In addition, oil and grease can block soil interstices, interfering with the ability of roots to draw water (Scott et al. 2004). Chlorine in particular can cause leaf tip burn. Pollutants, especially heavy metals, harm by accumulating in the soil over long periods of time, but they can also harm directly through irrigation (Hussain et al. 2002). Agronomic field experiments confirm reduced yields and crop quality from irrigation with industrially polluted water. Experiments have found rice to have more damaged grains and disagreeable taste, wheat to have lower protein content, and in general, plant height, leaf area, and dry matter to be reduced (World Bank and State Environmental Protection Administration 2007).

A few small case studies suggest that the findings of these field experiments extend to real-world settings. Reddy and Behera (2006) found an 88% decline in cultivated area in a village immediately downstream of an industrial cluster in Andhra Pradesh, India. Lindhjem et al. (2007) found that farmland irrigated with wastewater had lower corn and wheat production quantity and quality in Shijiazhuang, Hebei Province, China. Khai and Yabe (2013) found that areas in Can Tho, Vietnam irrigated with industrially polluted water had 12 percent lower yields and 26 percent lower profits. History also suggests that crop loss from industrial water pollution is not unknown to farmers; Patancheru, Andhra Pradesh saw massive farmer protests and a grassroots lawsuit in the late 1980s (Murty and Kumar 2011).

In contrast with industrial wasetewater, domestic or municipal wastewater can sometimes have positive effects on crop growth due to the nutrient value (Hussain et al. 2002). This is especially true for treated municipal wastewater. However, undiluted untreated wastewater can in fact have levels of nitrogen, phosphorous, and potassium that are so high they harm crop growth, and it poses health risks to agricultural workers, potentially reducing labor supply.

2.2 Welfare measures

I estimate the effects of industrial water pollution on agricultural revenue. Revenue is not directly interpretable as welfare, but I argue it is a reasonable proxy given conceptual and data limitations.

One candidate measure of rural welfare is agricultural profits: revenues net of costs. While some cost data is available, unfortunately it tends to be noisy and incomplete: for example, it fails to include the value of household labor. Estimating lost revenue will underestimate lost profits if farmers apply additional compensatory inputs, and it will overestimate lost welfare if farmers migrate or substitute other factors of production away from agriculture. However, I can measure many of these shifting factors, and I show these in one set of results.

A second candidate measure of welfare is agricultural yields as measured in crop quantities, taking the view that price fluctuations in both agricultural inputs and outputs are largely just transfers between consumers and producers, and that the welfare of both matters. The claim is that, in the case of additional inputs, lost profits overestimates total lost welfare because, for example, purchasing more fertilizer increases the welfare of fertilizer producers. However, this is also incomplete, failing to consider the opportunity cost of input factors of production and the fact that consumer substitution from any single crop is positively elastic. Nevertheless, I also report estimates on agricultural yields in one set of results.

A final candidate measure of welfare is land values or land rental rates. Unfortunately, systematic data do not exist on these with enough spatial coverage for my study. Even if they did, property values are a highly imperfect welfare measure in the presence of location preference heterogeneity, moving costs, or other imperfections in the land market which are likely high in India (Moretti 2011).

3 Data

3.1 Sources

3.1.1 Comprehensive Environmental Performance Index (CEPI)

The Central Pollution Control Board (CPCB) selected 88 industrial sites for detailed, long-term study in 2009. Names of these sites were taken from the CPCB document *Comprehensive Environmental Assessment of Industrial Clusters* (2009a). I identified the geolocation of each site using Google Earth and other publicly available reference information. These sites are displayed as red dots in Figure 1.

The CPCB document also contains numerical scores for air, water, and land pollution, and an overall score, each out of 100. Details of the scoring methodology are provided in the companion document *Criteria for Comprehensive Environmental Assessment of Industrial Clusters* (2009b). The CPCB considers a site "severely polluted" if the score for a single pollution type exceeds 50, or if the overall score exceeds 60 (the overall score is a nonlinear combination of the component scores). Sixty-three sites received such a "severe" rating in water pollution in 2009; these constitute my sample.

3.1.2 Pollution data

I use a rich dataset of water pollution measurements along rivers in India collected by the Central Pollution Control Board, as collected and made publicly available by Greenstone and Hanna (2014). This dataset includes 55,461 observations from 459 monitoring stations along 145 rivers between the years 1986 and 2005. The CPCB's water quality monitoring network has been expanding rapidly and presently there are over 2,500 monitoring stations. In future work I plan to incorporate this additional data since 2005.

The Greenstone & Hanna data include a noisy location measure as well as river name and a decsription of the sampling location. I identified, refined, or corrected the geolocation of each station using these variables, Google Earth, other CPCB documents, and other publicly available reference information. The locations of these stations are displayed as green dots in Figure 1.

Many water quality parameters have been collected by the CPCB at some point. However, only a few parameters are measured consistently. I focus on chemical oxygen demand (COD), a standardized laboratory test that serves as an omnibus meaure of organic compounds, which factories typically generate in high quantities. Along with the related but narrower test of biochemical oxygen demand (BOD), COD is the Indian government's top priority in regulating industrial wastewater (Duflo et al. 2013). One disadvantage of focusing on COD is that it misses inorganic pollutants like heavy metals, which, physiologically, are leading candidates for harming crop growth. Unfortunately, heavy metals are measured in less than three percent of observations. Therefore my best option is to use COD to proxy for all industrial pollution.

Another disadvantage to COD is that it is not an exclusive meausure of industrial pollution; it also can indicate the presence of domestic or municipal pollution (i.e., untreated sewage).

Because many industrial sites are located in cities, they may be coincident with domestic pollution, confounding the interpretation of my results as being driven by industrial pollution. However, another consistently measured water quality indicator is fecal coliforms, a count of the number of certain types of bacteria that originate from human waste. Because fecal coliforms are overwhelmingly produced by domestic pollution, not industrial pollution, partialing out fecal coliforms from COD should leave only the component of COD that is *not* produced by domestic pollution, i.e., industrial pollution. Therefore in some specifications I nonparametrically control for fecal coliforms.

3.1.3 Agricultural outcomes

For data on agricultural outcomes I use the Meso Scale Dataset from ICRISAT's Village Dynamics in South Asia project. This is a comprehensive district-level panel from 1966 to 2009. My primary outcome of interest is agricultural revenue. To calculate this, I multiply the quantity of each of 16 crops available in the dataset by the mean price for that crop in that district between 1966 and 1971. For districts without price data, I impute the state mean if available or the national mean otherwise.

I also construct the following variables for use as baseline covariates and outcomes:

- Climate variables: Average precipitation (mm/year); average potential evapotranspiration (mm/year).

- Irrigation variables: Land irrigated (net irrigated area per total district area, %); crops irrigated (gross irrigated area per gross cropped area, %); surface water and groundwater specific measures of the two variables above; pump density (number of diesel and electric pumps per district area, in thousands of hectares).

- Other agricultural variables: Land cropped (net cropped area per total district area, %); fertilization intensity (total fertilizer consumed per district area, in thousands of hectares).

- Population variables: Rural population density (total rural population per district area, in thousands of hectares); farmer density (total cultivators and agricultural workers per district area, in thousands of hectares).

3.1.4 Census

For geographic district boundaries and some population data I use the historic GIS shapefile of India's 1971 Census, from ML Infomap. I match districts to the ICRISAT data on state and district name. The ICRISAT data uses district definitions from 1966, so the boundaries may be slightly different, but a GIS shapefile of 1966 district boundaries was not available, and district boundaries changed little between 1961 and 1971.

Using ArcGIS, I calculate the longitude and latitude of the centroid of each district, as well as the land area in square kilometers. I also use the following variables from the 1971 census, included in the shapefile, as baseline covariates: total population density (people per square kilometer); urbanization rate (%); scheduled castes (%); scheduled tribes (%); and literacy rate (%).

3.2 Matching pollution data to industrial sites

I match the pollution observations with polluted industrial sites via the following procedure. First, I obtain a digital elevation model (DEM) at 15 arc-second resolution for the South Asia area from the HydroSHEDS project of the United States Geological Survey. From this DEM, I use the Hydrology tools in ArcGIS to obtain three principal products:

1. A vectorized river network, consisting of river segments (edges) and junctions (nodes).

2. A Shreve stream order system. This numbers each segment of a river such that the number increases at every junction; i.e., two lower-numbered tributaries always combine to form a higher-numbered river, regardless of whether one tributary has the same name as the downstream river segment.

3. A flow length raster, which calculates from each cell the distance along rivers that a particle would travel before reaching an ocean (or the edge of the raster).

I snap each industrial site to the nearest river within 10 km (only a few do not meet this criteria), and for each industrial site and pollution obsevation station I extract the river segment, stream order, and flow length. I join each industrial site to all pollution stations on the same river network, and kept only pairs that meet criteria for being either upstream or downstream of each other (i.e., a site and station on separate tributaries are neither).

I construct distance, which I use as a running variable, as the difference in flow lengths between station and industrial site, and a downstream indicator variable equaling one if the distance variable is positive (station is downstream of industrial site). Finally, I assign each station to its nearest industrial site and discard other matches, so that each station is only in the final dataset once.

3.3 Matching agricultural data to industrial sites

I identify districts upstream and downstream of each industrial site by inspection in ArcGIS using district and river maps and elevation, flow direction, and flow length data. I first identify the district the industrial site lies in and classify it according to whether more of its land lies upstream or downstream of the site. I then select 1-2 districts upstream of this district and 1-2 districts downstream. The final sample consists of minimum 3 and maximum 5 districts per industrial site. In future work I will formalize this procedure and calculate the percentage of each district's land area which lies in the watershed (upstream) or catchment area (downstream) of each site.

3.4 Covariate balance

3.4.1 Full sample

Table 1 presents the balance of many covariates in 1971-72 across districts categorized as either upstream or downstream of industrial sites. The first column shows the mean and standard deviation of upstream districts, the second shows the same for downstream districts, and the third column shows the difference and the standard error of a t-test comparing the two means. For some variables whose distributions are very right-skewed, I take the natural logarithm of the data and show the same comparisons.

The two groups are somewhat similar. Downstream districts have slightly more scheduled castes and slightly fewer scheduled tribes, but since both are disadvantaged categories, socioeconomic characteristics should be similar. More worrying is that downstream districts have substantially more irrigation, especially by surface water. Irrigation is a very important factor in determining agricultural productivity, so this diference may confound the relationship between polluted factories and agricultural revenues or yields. The difference is likely due to dams which make it easier to irrigate downstream and not upstream. Currently I do not use data on dams; in future work I can directly control for them.

3.4.2 Balanced sample

In an attempt to construct a sample in which upstream and downstream districts are more similar at baseline, I discard industrial sites which have an upstream-downstream difference in either land irrigated or crops irrigated exceeding 20 percentage points. This is admittedly a fairly arbitrary threshold but one that represents a natural break in the histograms of these differences.

Table 2 presents again the covariate balance for this trimmed sample. Now the irrigation variables are extremely similar between upstream and downstream districts. In fact, none of the covariates has a statistically significant difference. I show results using both the full sample and this balanced sample.

4 Empirical strategy

4.1 Regression discontinuity in pollution

The spatial discontinuity in pollution and the linear flow of rivers presents a natural setting for a regression discontinuity design. I match each polluted industrial site to upstream and downstream pollution measurement stations on the same river path, and I line up these stations in river space such that each industrial site is at zero. The running variable is distance along a river path from the industrial site, with positive values indicating that a measurement station is downstream of an industrial site. I estimate local linear regressions of the following form:

$$cod_{sit} = \beta downstream_{si} + \gamma distance_{si} + \delta distance_{si} downstream_{si} + \alpha_{st} + \varepsilon_{sit} \tag{1}$$

where cod_{sit} is chemical oxygen demand measured at station i in year t, which spans the years 1986 to 2005. Variable $downstream_{si}$ is the discontinuous treatment; it takes a value of 1 if station i lies downstream of industrial site s and 0 if upstream. I estimate local linear regressions without higher order polynomials, following the advice of Gelman and Imbens (2014), allowing both the intercept and slope of the running variable $distance_{si}$ to vary on opposite sides of the discontinuity.

I include site-by-year fixed effects α_{st}, so that the treatment effect at the discontinuity is identified only using variation between upstream and downstream measurements on the same river path in the same year. I use a triangular kernel, which is optimal for estimating local linear regressions at a boundary (Fan and Gijbels 1996). I cluster standard errors by station and report results using a range of reasonable bandwidths.

For the regression discontinuity graph, I first partial out the state-by-year fixed effects from the COD measurements by running a triangular-weighted local linear regression of cod_{sit} on α_{st}. I then graph the residuals from this regression plus the upstream mean of the fixed effects, and fit local linear regressions of these residuals on $distance_{si}$, $downstream_{si}$, and their interaction.

4.2 Spatial matching in agricultural outcomes

Unfortunately, agricultural data is not available at a finer spatial resolution than district, making a regression discontinuity design infeasible for these outcomes. Instead, I simply match upstream and downstream districts, estimating ordinary least squares regressions of the following form:

$$y_{sit} = \beta downstream_{si} + \Pi X_{si} + \alpha_{st} + \varepsilon_{sit} \tag{2}$$

where y_{sit} is an agricultural outcome measured in district i in year t, X_{si} is a vector of baseline covariates and α_{st} are site-by-year fixed effects. The agricultural data spans 1966-2009; 1972 is the first year for which all covariates are observed, so I divide the time series there, including baseline covariates from 1972 and estimating the regressions only on data from 1973 onward. Site-by-year fixed effects again ensure that the treatment effect of being downstream is identified only using differences between upstream and downstream districts near the same industrial site in the same year. I cluster standard errors by district to account for serial correlation.

The identifying assumption is that no other factors that are relevant to agricultural outcomes also vary systematically between upstream and downstream districts. This is potentially plausible, since the polluted industrial sites are located all over India, on rivers oriented in all directions. This assumption does not seem to hold very well in the full sample (Table 1), but it does seem to hold on many observed factors in the balanced sample (Table 2). This of course does not ensure that it holds on unobserved factors, and so in future work I will combine this spatial difference with sources of time series variation, which will allow me to remove time-invariant district-specific characteristics.

5 Results

5.1 Pollution

I first show that the industrial sites considered "severely polluted" by the Central Pollution Control Board do in fact increase pollution levels discontinuously in nearby rivers.

Regression discontinuity. Figure 2 presents photographic evidence of this discontinuity at one site in my sample: the Nazafgarh Drain Basin on the Yamuna River just north of New Delhi. The river flows from north to south and enters the image at the top with a green color. In the center of the image, an industrial effluent channel meets the river, discontinuously turning the river black. Color is neither a sufficient nor necessary condition for industrial water pollution, but it confirms the presence of water from another source and is correlated with pollution.

Moving to the full sample and using direct, quantitative measurements of pollution, Figure 3 provides visual evidence of the discontinuity in pollution concentrations at polluting industrial sites. This figure plots measurements of chemical oxygen demand (COD) - an omnibus measure of water pollution - after partialing out site-by-year fixed effects, against distance from a polluting industrial site. Positive distance values (right of the vertical line) indicate that the measurement is downstream; negative values (left of the vertical line) are upstream. Data is collapsed to 3-km wide bins; the light blue circles plot the mean within each bin, with the diameter of the circle proportional to the number of observations falling within that bin. The dark blue lines represent a local linear regression at $distance = 0$ using a bandwidth of 50 km and a triangular kernel, in which both intercept and slope are allowed to vary on opposite sides of $distance = 0$.

Visually there appears to be evidence of a large, discontinuous rise in pollution concentrations at $distance = 0$, coinciding with the presence of large polluting industrial sites. This confirms that the industrial sites selected by Central Pollution Control Board for detailed study are indeed releasing large amounts of water pollution into India's river systems. Pollution levels appear to fall downstream of the site; this makes sense because the further downstream the river runs, the more that the original pollution concentrations in the river are diluted with runoff from other areas. It is less clear why pollution might rise approaching the site from upstream; however, the estimated upstream slope is sensitive to bandwidth and often smaller or negative.

Varying bandwidths. Table 3 makes this visual evidence more formal. Panel A presents results of eight local linear regressions of the form 1, with bandwidth varying from 200 km to 7 km, the smallest bandwidth for which I can estimate all parameters of the regression. The first row shows the coefficient on the *downstream* indicator, which is the estimate of the average discontinuity in pollution at an industrial site. For all bandwidths this estimate is positive, large, and has a 99% or 95% confidence interval that excludes zero. Downstream pollution concentrations are expected to fall with increasing distance from a pollution source; accordingly, the estimate of the discontinuity grows larger in magnitude as the bandwidth narrows and we zoom in on the polluting industrial site.

At the third-smallest bandwidth, 15 km, the discontinuity at an industrial site is 95 mg/L. Comparing this to the upstream mean (the estimate of the limit of COD concentrations as *distance* approaches 0 from the upstream side) of 34 mg/L, this implies that the average "severely polluted" industrial site nearly triples pollution levels in the nearest river.

Fecal coliforms. Panel B of Table 3 shows the same local linear regressions but linearly controlling for fecal coliforms. The distribution of this variable is heavily skewed right so I use its natural logarithm and allow its slope to differ upstream and downstream of the industrial sites. While COD is produced by both industrial and domestic sources, fecal coliforms are overwhelmingly produced by domestic sources only, so by controlling for the latter, the remaining variation in COD should reflect only industrial sources. The results here are quite consistent with Panel A. As expected, the estimates of the pollution discontinuity are smaller, but still large, positive, and statistically significant at bandwidths smaller than 50 km. This suggests that the discontinuous rise in pollution at the point of "severely polluted" industrial sites does in fact reflect industrially generated pollution, not just domestic sources which happen to coincide with the industrial sites.

5.2 Agricultural outcomes

Having shown that water pollution in rivers rises discontinuously at "severely polluted" industrial sites, I investigate the downstream effects this pollution has on agriculture, using the sample trimmed by excluding industrial sites with a large upstream-downstream difference in baseline irrigation rates.

Revenues. Table 4 reports the results of regressions of the natural logarithm of total agricultural revenues on an indicator for being downstream, baseline covariates, and site-by-year fixed effects, according to equation 2. The sample is limited to 3-5 districts immediately surrounding each polluting industrial site, and the fixed effects ensure the identifying variation is only between upstream and downstream districts around the same site in the same year.

The leftmost column includes no baseline covariates, the middle columns include one group of covariates each, and the rightmost column includes all baseline covariates. Because the dependent variable is in natural logarithms, the coefficients can be interpreted as roughly proportional changes in revenues. The point estimates of the effect of being downstream are almost all negative, and I cannot statistically reject that they are all equal. When all baseline covariates are included, being downstream is associated with a 9 percent reduction in agricultural revenues, with a 95% confidence interval of (-2, -17) log points.

Other outcomes. Table 5 shows the results of similar regressions for a variety of other agricultural outcome variables, each including all baseline covariates. These estimates can provide evidence on the channels through which the negative effect on revenues operates, or alternatively whether farmers may be engaging in compensatory behavior, increasing agricultural inputs to soften the impact on revenues but at a higher cost. Column (1) is repeated from Table 4, for ease of comparison.

The evidence suggests that the entire effect of industrial water pollution operates through reduced yields, i.e., revenues per cropped area, and not through reallocation of factors of production such as land, labor, irrigation, or fertilizer away from agriculture. Moreover, farmers do not appear to be compensating for the reduced yields by applying additional inputs. Of course, it is possible that there are heterogeneous effects: some farmers respond to water pollution by substituting factors away from farming, while others maintain revenue by applying more inputs, but on average there is no evidence that either is happening systematically.

This can be seen by noting that the effect of being downstream on yields is 16% and even more precisely estimated than the effect on revenues, while the effects on all other outcomes are smaller and most not significantly different from zero. Land area irrigated and crop area irrigated appear to decline by 2%, but their 95% confidence intervals do not exclude zero in the balanced sample, and these changes could explain at most 20% of the effect on revenues. The effect of being downstream of polluted industrial sites on rural population and numbers of farmers are imprecise but the point estimates are positive, suggesting that the fall in revenues is not coming from out-migration or exit from farming.

6 Conclusion

This paper provides the first quasi-experimental evidence on the costs of industrial water pollution to agriculture. I examine 63 industrial sites in India identified by the government as "severely polluting" and estimate the costs of their pollution to downstream agriculture. First, I find that the location of these industrial sites coincides with a large, discontinuous jump in water pollution in nearby rivers. Second, I find that each district immediately downstream of these sites has, on average, 9 percent lower yields (with a 95% confidence interval of -2% to -17%) than a corresponding district immediately upstream of the same site, in the same year. Third, I find that this effect is driven by the direct impact on yields; there is no evidence that factor reallocation either mitigates or exacerbates it.

In future work I plan three major improvements to enhance the credulity of these estimates. First, I plan to use a high-resolution vegetation index derived from NASA satellite data to improve the spatial resolution of the agricultural outcomes. Currently the district-to-district comparison is rather coarse; remote sensing data will permit me to implement a true spatial regression discontinuity design on a proxy for crop yields. Second, I plan to combine my existing spatial variation with temporal variation in the number of polluting factories in each industrial site or corresponding district. Under my current empirical strategy, there is still a possibility that upstream and downstream districts differ systematically in unobserved ways that affect agricultural outcomes, confounding the effect of industrial pollution. Adding temporal variation will allow me to control for unobserved time-invariant factors.

Third, I plan to also incorporate short-term temporal variation from a river flow instrument. This instrument is inverse flow volume as predicted by upstream rainfall. Economic outcomes depend on pollution concentrations, the ratio of pollution to water volume; upstream rainfall increases the denominator without affecting the numerator. By comparing these results to those using the number of polluting factories over time, I can provide evidence on the difference between short- and long-term effects of pollution.

References

Adhvaryu, Achyuta, Namrata Kala, and Anant Nyshadham, “Management and Shocks to Worker Productivity,” 2016.

Aragón, Fernando M. and Juan Pablo Rud, “Polluting Industries and Agricultural Productivity: Evidence from Mining in Ghana,” *The Economic Journal*, 2016, *126* (597), 1980–2011.

Arceo, Eva, Rema Hanna, and Paulina Oliva, “Does the Effect of Pollution on Infant Mortality Differ Between Developing and Developed Countries? Evidence from Mexico City,” *The Economic Journal*, 2016, *126* (591), 257–280.

Brainerd, Elizabeth and Nidhiya Menon, “Seasonal effects of water quality: The hidden costs of the Green Revolution to infant and child health in India,” *Journal of Development Economics*, 2014, *107*, 49–64.

Bustos, Paula, Bruno Caprettini, and Jacopo Ponticelli, “Agricultural Productivity and Structural Transformation: Evidence from Brazil,” *American Economic Review*, 2016, *106* (6), 1320–1365.

Central Pollution Control Board, “Comprehensive Environmental Assessment of Industrial Clusters,” Technical Report December 2009.

_ , “Criteria for Comprehensive Environmental Assessment of Industrial Clusters,” Technical Report December 2009.

Chen, Yuyu, Avraham Ebenstein, Michael Greenstone, and Hongbin Li, “Evidence on the impact of sustained exposure to air pollution on life expectancy from China’s Huai River policy.,” *Proceedings of the National Academy of Sciences of the United States of America*, aug 2013, *110* (32), 12936–12941.

Do, Quy Toan, Shareen Joshi, and Samuel Stolper, “Can environmental policy reduce infant mortality? Evidence from the Ganga Pollution Cases,” *Journal of Development Economics*, 2018, *133* (September 2016), 306–325.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan, “Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India,” *The Quarterly Journal of Economics*, 2013, pp. 1–47.

Fan, Jianqing and Irene Gijbels, “Local Polynomial Modelling and its Applications,” *Monographs on Statistics and Applied Probability*, 1996, *66.*

Gelman, Andrew and Guido Imbens, “Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs,” *National Bureau of Economic Research Working Paper Series*, 2014, *No. 20405.*

Ghatak, Maitreesh and Dilip Mookherjee, “Land acquisition for industrialization and compensation of displaced farmers,” *Journal of Development Economics*, 2014, *110*, 303–312.

Greenstone, Michael and Rema Hanna, “Environmental Regulations, Air and Water Pollution, and Infant Mortality in India,” *American Economic Review*, oct 2014, *104* (10), 3038–3072.

Hussain, Intizar, Liqa Raschid, Munir A. Hanjra, Fuard Marikar, and Wim van der Hoek, *Wastewater Use in Agriculture: Review of Impacts and Methodological Issues in Valuing Impacts* 2002.

Jayachandran, Seema, "Air Quality and Early-Life Mortality Evidence from Indonesia's Wildfires," *Journal of Human Resources*, 2009, *44* (4).

Keiser, David A. and Joseph S. Shapiro, "Consequences of the Clean Water Act and the Demand for Water Quality," 2017.

Khai, Huynh Viet and Mitsuyasu Yabe, "Impact of Industrial Water Pollution on Rice Production in Vietnam," in "International Perspectives on Water Quality Management and Pollutant Control" 2013.

Lindhjem, Henrik, Tao Hu, Zhong Ma, John Magne Skjelvik, Guojun Song, Haakon Vennemo, Jian Wu, and Shiqiu Zhang, "Environmental economic impact assessment in China: Problems and prospects," *Environmental Impact Assessment Review*, 2007, *27* (1), 1–25.

Moretti, Enrico, *Local Labor Markets*, Vol. 4, Elsevier B.V., 2011.

Murty, M.N. and Surender Kumar, "Water Pollution in India: An Economic Appraisal," in "India Infrastructure Report" 2011.

Reddy, V. Ratna and Bhagirath Behera, "Impact of water pollution on rural communities: An economic analysis," *Ecological Economics*, 2006, *58* (3), 520–537.

Scott, C.I., N. I. Faruqui, and L. Raschid-Sally, *Wastewater Use in Irrigated Agriculture: Confronting the Livelihood and Environmental Realities* 2004.

World Bank and State Environmental Protection Administration, "Cost of Pollution in China: Economic Estimates of Physical Damages," Technical Report 10 2007.

Table 1: Covariate balance (full sample).

	Levels			**Logs**		
	Up	Down	Diff	Up	Down	Diff
Latitude (degrees)	23.4	24.5	1.1			
	[5.1]	[4.8]	(0.8)			
Longitude (degrees)	79.3	78.6	-0.7			
	[4.0]	[3.8]	(0.6)			
Land area (km2)	9951	8787	-1165			
	[5497]	[5921]	(949)			
Population density, 1971 (people/km2)	259	282	23			
	[157]	[142]	(25)			
Urban, 1971 (%)	0.21	0.19	-0.01			
	[0.13]	[0.12]	(0.02)			
Scheduled castes, 1971 (%)	0.14	0.17	0.03***			
	[0.06]	[0.06]	(0.01)			
Scheduled tribes, 1971 (%)	0.11	0.05	-0.06**			
	[0.19]	[0.10]	(0.02)			
Literacy rate, 1971 (%)	0.27	0.27	-0.01			
	[0.08]	[0.08]	(0.01)			
Avg. precipitation (mm/yr)	1027	934	-93			
	[397]	[324]	(59)			
Avg. potential evapotranspiration (mm/yr)	1579	1569	-10			
	[200]	[148]	(29)			
Land cropped, 1972 (%)	0.56	0.62	0.07**			
	[0.21]	[0.17]	(0.03)			
Land irrigated, 1972 (%)	0.19	0.27	0.09**	-2.48	-1.84	0.64***
	[0.20]	[0.22]	(0.04)	[1.47]	[1.27]	(0.23)
Crops irrigated, 1972 (%)	0.28	0.38	0.10**	-1.82	-1.33	0.49***
	[0.24]	[0.25]	(0.04)	[1.18]	[1.01]	(0.18)
Land irrigated by surface water, 1972 (%)	0.07	0.12	0.05***	-3.89	-2.95	0.94***
	[0.096]	[0.11]	(0.02)	[1.83]	[1.83]	(0.30)
Land irrigated by groundwater, 1972 (%)	0.11	0.15	0.03	-3.36	-2.85	0.50
	[0.13]	[0.14]	(0.02)	[1.89]	[1.80]	(0.31)
Crops irrigated by surface water, 1972 (%)	0.11	0.18	0.07***	-3.21	-2.43	0.78***
	[0.13]	[0.15]	(0.02)	[1.69]	[1.75]	(0.29)
Crops irrigated by groundwater, 1972 (%)	0.16	0.20	0.04	-2.68	-2.33	0.35
	[0.17]	[0.18]	(0.03)	[1.58]	[1.53]	(0.26)
Water pumps per land area, 1972 (km-2)	0.019	0.021	0.002	-4.94	-4.76	0.18
	[0.026]	[0.021]	(0.004)	[1.69]	[1.80]	(0.29)
Fertilizer per cropped area, 1972 (ton/Kha)	19.3	21.3	2.0	2.45	2.58	0.13
	[17.2]	[16.3]	(2.8)	[1.17]	[1.54]	(0.23)
Observations	67	80				

Standard deviations in brackets. Standard errors of t-tests in parentheses.

Figure 1: Locations of "severely polluted" industrial sites (red dots) and water pollution measurement stations (green dots).

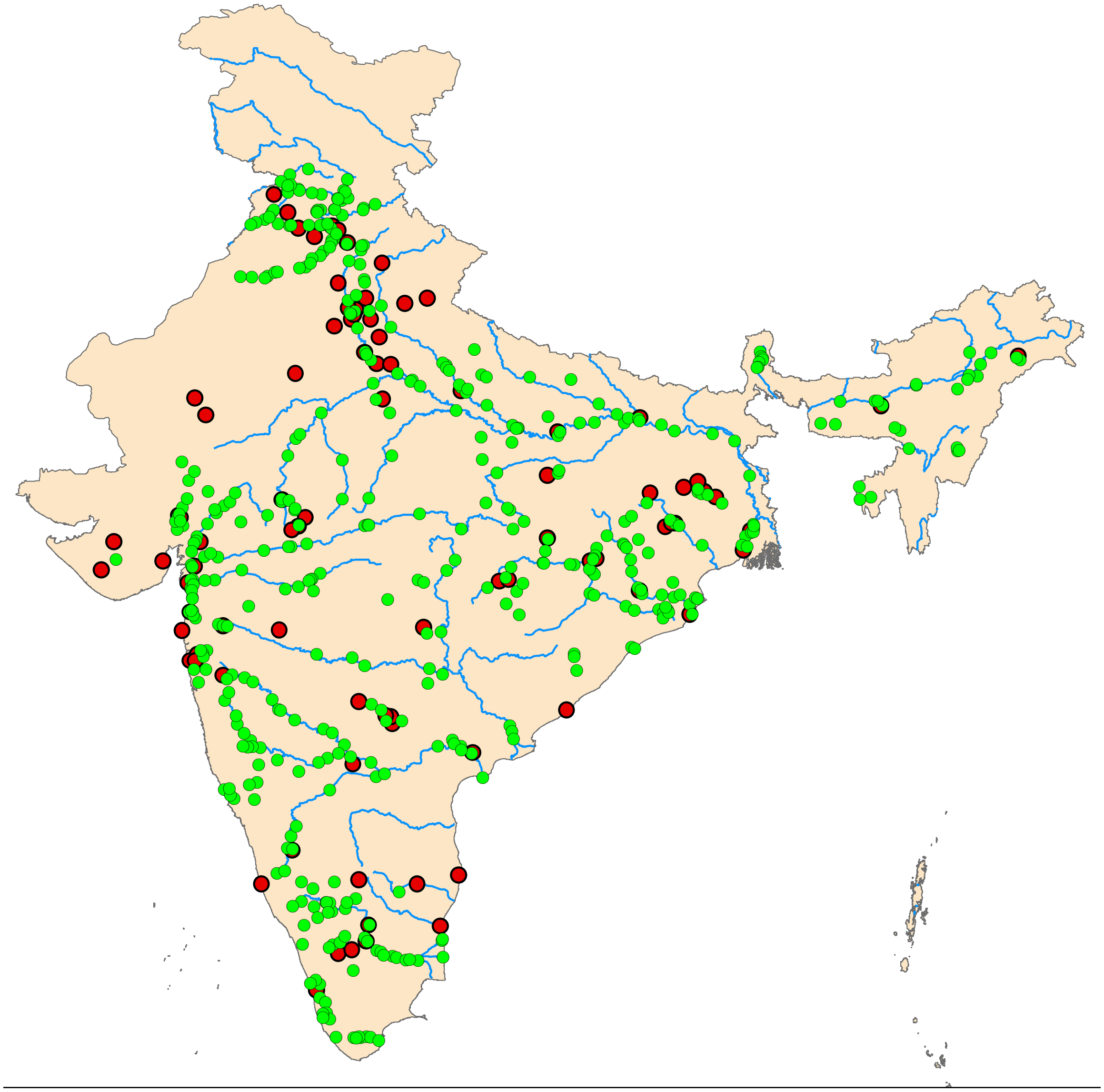

Table 2: Covariate balance (sample trimmed by excluding sites with a large upstream-downstream difference in baseline irrigation rates).

	Levels			**Logs**		
	Up	Down	Diff	Up	Down	Diff
Latitude (degrees)	23.4	24.0	0.6			
	[5.1]	[4.6]	(1.0)			
Longitude (degrees)	77.8	78.1	0.3			
	[3.2]	[3.6]	(0.7)			
Land area (km2)	9468	9258	-210			
	[5328]	[6488]	(1203)			
Population density, 1971 (people/km2)	249	273	24			
	[142]	[144]	(29)			
Urban, 1971 (%)	0.21	0.19	-0.02			
	[0.13]	[0.13]	(0.03)			
Scheduled castes, 1971 (%)	0.15	0.16	0.02			
	[0.05]	[0.06]	(0.011)			
Scheduled tribes, 1971 (%)	0.06	0.05	-0.01			
	[0.13]	[0.11]	(0.02)			
Literacy rate, 1971 (%)	0.28	0.27	-0.02			
	[0.08]	[0.09]	(0.02)			
Avg. precipitation (mm/yr)	911	906	-5			
	[371]	[294]	(66)			
Avg. potential evapotranspiration (mm/yr)	1622	1589	-33			
	[200]	[153]	(35)			
Land cropped, 1972 (%)	0.63	0.63	0.002			
	[0.17]	[0.16]	(0.03)			
Land irrigated, 1972 (%)	0.22	0.24	0.02	-2.21	-2.02	0.19
	[0.21]	[0.21]	(0.04)	[1.42]	[1.30]	(0.27)
Crops irrigated, 1972 (%)	0.29	0.32	0.03	-1.73	-1.52	0.21
	[0.23]	[0.23]	(0.05)	[1.20]	[1.05]	(0.22)
Land irrigated by surface water, 1972 (%)	0.08	0.10	0.02	-3.77	-3.21	0.55
	[0.10]	[0.091]	(0.02)	[1.89]	[1.98]	(0.39)
Land irrigated by groundwater, 1972 (%)	0.14	0.13	0.00	-2.86	-3.03	-0.18
	[0.13]	[0.14]	(0.03)	[1.65]	[1.87]	(0.36)
Crops irrigated by surface water, 1972 (%)	0.12	0.15	0.03	-3.25	-2.71	0.54
	[0.13]	[0.13]	(0.03)	[1.78]	[1.90]	(0.37)
Crops irrigated by groundwater, 1972 (%)	0.19	0.18	-0.01	-2.34	-2.53	-0.19
	[0.17]	[0.17]	(0.04)	[1.45]	[1.62]	(0.31)
Water pumps per land area, 1972 (km-2)	0.021	0.017	-0.004	-4.53	-4.84	-0.31
	[0.024]	[0.015]	(0.004)	[1.38]	[1.67]	(0.31)
Fertilizer per cropped area, 1972 (ton/Kha)	20.2	17.7	-2.5	2.54	2.36	-0.18
	[16.8]	[13.5]	(3.0)	[1.16]	[1.69]	(0.30)
Observations	44	58				

Standard deviations in brackets. Standard errors of t-tests in parentheses.

Figure 2: Satellite photo of the discontinuity in river color at the outlet of the Nazafgarh Drain Basin on the Yamuna River, just north of New Delhi.

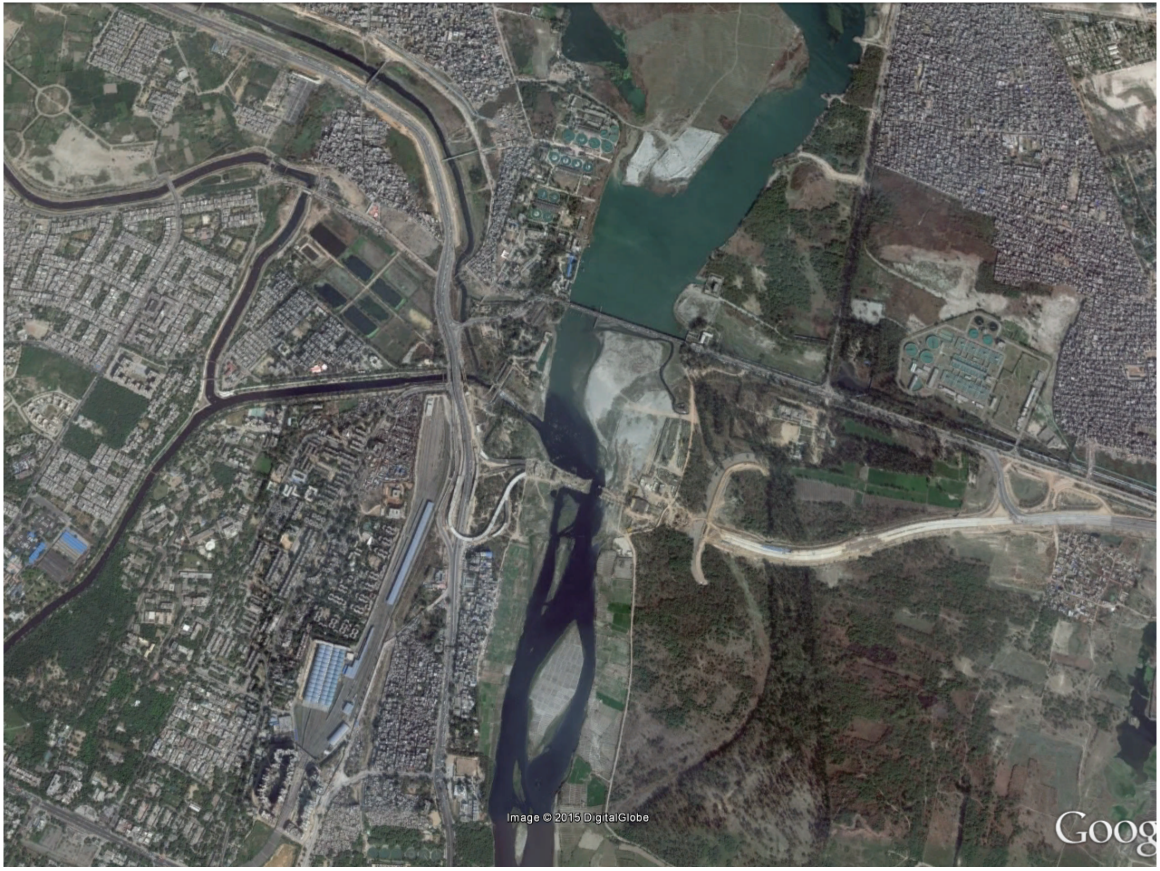

Figure 3: Discontinuity in river pollution levels at the point of “severely polluted” industrial sites.

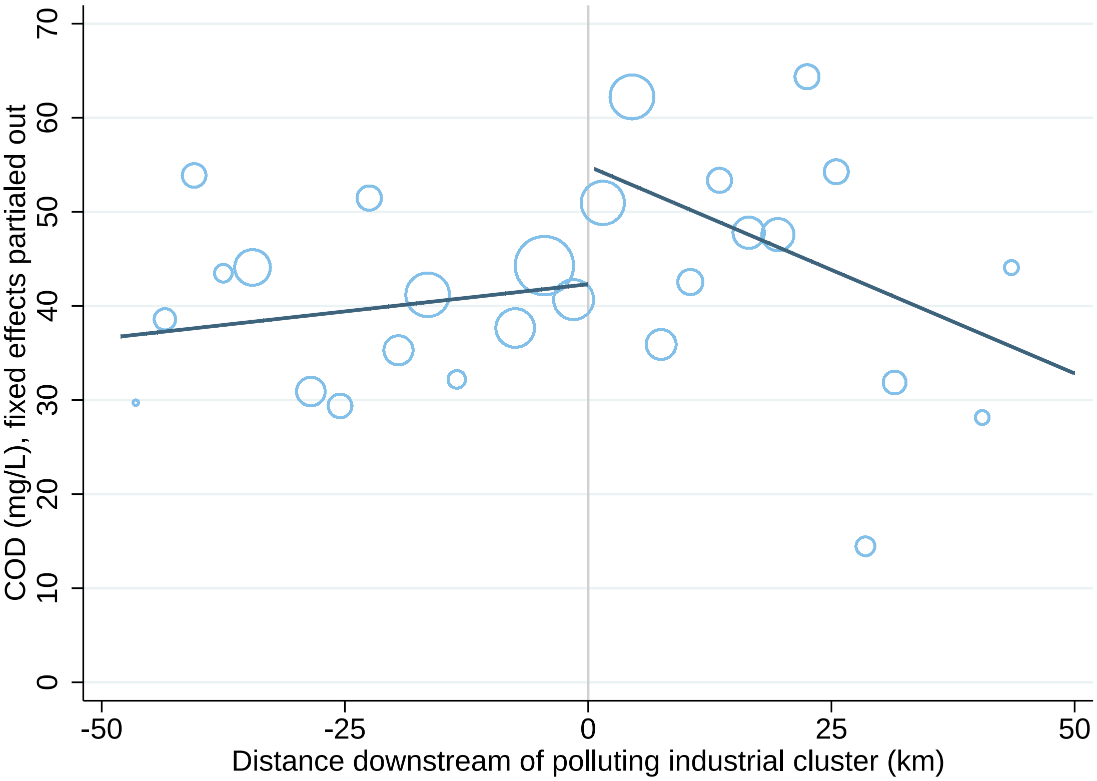

Table 3: Local linear regressions of water pollution estimated at the point of a "severely polluted" industrial site.

Dependent Variable: Chemical oxygen demand (COD), mg/L

Panel A: Benchmark RD

Bandwidth (km)	200	100	50	30	20	15	10	7
Downstream indicator	**12.6****	**19.7*****	**30.8*****	**38.3*****	**58.4*****	**95.0*****	**112.2*****	**152.4*****
	(5.6)	**(7.1)**	**(9.7)**	**(12.4)**	**(14.8)**	**(19.7)**	**(23.0)**	**(3.6)**
Distance along river (km)	0.1*	0.1	0.0	-0.2	-0.9	-2.6	-4.9	-11.0***
	(0.03)	(0.1)	(0.2)	(0.4)	(0.9)	(2.5)	(3.7)	(0.18)
Downstream X Distance	-0.3***	-0.6***	-1.0**	-0.9	-1.1	-2.3	0.0	5.8***
	(0.1)	(0.2)	(0.4)	(0.6)	(1.3)	(2.4)	(3.8)	(1.0)
Site-by-year fixed effects	X	X	X	X	X	X	X	X
Observations	13,767	10,150	7,437	6,282	5,002	3,715	3,234	2,744
Clusters (stations)	132	100	70	55	43	33	28	24
Upstream mean	36.8	40.3	40.4	40.0	39.9	34.2	28.1	15.1

Panel B: Partialing out Fecal Coliforms

Bandwidth (km)	200	100	50	30	20	15	10	7
Downstream indicator	**5.0**	**13.8**	**19.6***	**28.8****	**44.7*****	**82.4*****	**95.0*****	**111.2*****
	(8.4)	**(8.3)**	**(10.2)**	**(11.9)**	**(14.2)**	**(22.1)**	**(17.7)**	**(12.1)**
Distance along river (km)	0.0	0.1	0.0	-0.2	-0.8	-3.5	-6.9**	-9.4***
	(0.0)	(0.1)	(0.1)	(0.3)	(0.7)	(2.5)	(2.5)	(0.6)
Downstream X Distance	-0.2***	-0.4***	-0.5	-0.2	-0.4	0.2	6.5**	10.4***
	(0.1)	(0.1)	(0.3)	(0.4)	(1.0)	(2.4)	(2.8)	(2.7)
Ln Fecal coliforms	2.5***	2.7***	2.0**	1.9*	1.6	1.3	1.1	2.1
	(0.7)	(0.8)	(0.9)	(1.0)	(1.2)	(1.4)	(1.5)	(1.7)
Downstream X Ln Fecal Coli.	0.7	0.2	0.5	-0.2	0.0	0.2	0.7	-0.4
	(1.0)	(0.9)	(0.9)	(1.1)	(1.5)	(1.9)	(2.3)	(2.6)
Site-by-year fixed effects	X	X	X	X	X	X	X	X
Observations	10,577	7,530	5,297	4,381	3,423	2,577	2,175	1,804
Clusters (stations)	129	97	68	53	42	31	26	22
Upstream mean	19.4	21.0	25.4	25.2	26.8	19.3	14.9	6.4

Standard errors clustered by station. * p<0.10, ** p<0.05, *** p<0.01

Table 4: Regressions of agricultural revenues comparing districts immediately upstream and downstream of "severely polluted" industrial sites.

Dependent Variable: Ln Total Agricultural Revenues (per total land area)

Downstream indicator	-0.01	-0.03	-0.10	-0.10	-0.03	0.01	-0.09**
	(0.11)	(0.13)	(0.09)	(0.10)	(0.05)	(0.07)	(0.04)
Geography (Lat/Lon)		X					X
Climate (Rain/PET)			X				X
Social (Dens/Urban/SC/ST)				X			X
Irrigation (%Land/%Crops/Surface/Ground)					X		X
Fertilizer						X	X
Site-by-year fixed effects	X	X	X	X	X	X	X
Observations	2,303	2,303	2,231	2,303	2,303	2,231	2,231
Clusters (districts)	53	53	51	53	53	51	51

Standard errors clustered by district. * p<0.10, ** p<0.05, *** p<0.01

Table 5: Regressions of several agricultural outcomes comparing districts immediately upstream and downstream of “severely polluted” industrial sites.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Agricultural Revenues	Yields	Cropped Area	Irrigated Area	Crops Irrigated	Fertilizer	Rural Population	Farmers
Downstream indicator	-0.09**	-0.16***	0.02	-0.02*	-0.02	-0.01	0.09	0.08
	(0.04)	(0.03)	(0.01)	(0.01)	(0.01)	(0.05)	(0.06)	(0.05)
Geographic covariates	X	X	X	X	X	X	X	X
Climate covariates	X	X	X	X	X	X	X	X
Population covariates	X	X	X	X	X	X	X	X
Irrigation covariates	X	X	X	X	X	X	X	X
Site-by-year fixed effects	X	X	X	X	X	X	X	X
Observations	2,231	2,249	2,237	2,174	2,170	2,227	180	177

Standard errors clustered by district.
* p<0.10, ** p<0.05, *** p<0.01

Dependent variable definitions:
(1) Ln Revenues (revenue per total land area) (INR/1000 hectares)
(2) Ln Agricultural Yields (revenue per cropped area) (INR/1000 hectares)
(3) Land cropped (%)
(4) Land irrigated (%)
(5) Crops Irrigated (%)
(6) Ln Fertilizer Consumption per 1000 hectares land area
(7) Ln Rural Population per 1000 hectares land area
(8) Ln Cultivators + Agricultural Workers per 1000 hectares land area