Industrial Water Pollution and Agricultural Production in India

Nick Hagerty
Anshuman Tiwari*

December 24, 2022

Abstract

Industrial water pollution is high in many developing countries but often receives less attention than air and domestic water pollution. We estimate the costs of industrial water pollution to agriculture in India, focusing on 48 industrial sites identified by the central government as "severely polluted." We exploit the spatial discontinuity in pollution concentrations that these sites generate along a river. First, we show that these sites do coincide with a large, sudden rise in pollutant concentrations in the nearest river. Then, we find that a remote sensing measure of crop yields is no lower in villages immediately downstream of polluting sites, relative to villages upstream of the same site in the same year. Downstream farmers switch irrigation sources from rivers and canals to wells in some specifications, suggesting costly input substitution may avert pollution damages. Damages to agriculture may not represent a major cost of water pollution, though many other social costs are not yet quantified.

*Hagerty: Montana State University (email: nicholas.hagerty@montana.edu); Tiwari: University of California, Santa Barbara and Environmental Defense Fund (email: atiwari@ucsb.edu). This paper supersedes an earlier working paper titled "The Costs of Industrial Water Pollution in Agriculture in India." We thank Sambath Jayapregasham for excellent research assistance. For helpful discussion and comments, we thank (without implicating) Abhijit Banerjee, Marshall Burke, Esther Duflo, Eyal Frank, Peter Hull, Simon Jäger, Peiley Lau, David Molitor, Alex Oberg, Ben Olken, Sheila Olmstead, Molly Sears, and Anant Sudarshan. Anshuman Tiwari gratefully acknowledges financial support from the Grantham Research Foundation.

1 Introduction

Pollution levels in low- and middle-income countries are often orders of magnitude worse than in high-income countries. Simple linear extrapolation suggests the costs to health, productivity, and ecology could be high – and they could be even higher if they are nonlinear, as some evidence suggests, with marginal costs increasing in pollution levels (Arceo et al. 2016). Unfortunately, most causal evidence on the costs of pollution comes from developed countries, with little basis to extrapolate to developing settings. Water pollution in particular has received less attention from both researchers and the public than air pollution. In India, while regulation on air pollution may have reduced some air pollutants due to public pressure, similarly strict regulation has not discernibly improved water quality (Greenstone and Hanna 2014). Toxic white foam now forms annually on water bodies in New Delhi and Bengaluru (Möller-Gulland 2018), and mass fish deaths have become common (Vyas 2022).

Even in high-income countries, the social costs of water pollution have been challenging to quantify. While surveys show high levels of public interest in water quality, research has rarely found economically significant impacts of water pollution. This could be because the costs truly are low, or alternatively because water pollution is especially difficult to study. Low quality and availability of pollution measurements, the difficulty of modeling complex spatial relationships, and the wide variety of distinct pollutants may have both inhibited research and attenuated estimates that do exist (Keiser and Shapiro 2019b).

This paper estimates the effects of industrial water pollution on agricultural production in India. We study agriculture because several reasons suggest it could be the site of large aggregate effects of water pollution. Agriculture uses four times more water than all other sectors of the economy combined (FAO 2018), and irrigation water is rarely treated before use, unlike drinking water. The agricultural sector is also large and ubiquitous, so it can be found near virtually every source of pollution. We focus on 48 industrial sites identified by India's Central Pollution Control Board in 2009 as "severely polluted" with respect to water pollution. India's industrial clusters are home to some of the greatest concentrations of industrial pollution in the world (Mohan 2021), so if

industrial water pollution matters anywhere, it likely matters here.

Our research design exploits the fact that water pollution, unlike air pollution, almost always flows in only one direction from its source. 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. We use hydrological modeling to precisely determine areas that are upstream and downstream and compute spatial relationships.

Importantly, we measure the overall effect of high-polluting industrial sites, rather than specific pollutants. This approach allows us to sidestep the need to rely on water quality monitoring data, which are generally plagued by noise, infrequency, low spatial density, and site selection bias. They are also difficult to summarize, since industrial effluents can contain thousands of distinct elements and compounds. Any of these could independently harm human, crop, or ecosystem health, but each typically requires a separate laboratory test to measure.

To measure agricultural outcomes, our main analysis relies on satellite data. No other data source is available at high enough spatial resolution to enable a spatial regression discontinuity design; even in the United States, aggregate statistics are too coarse and agricultural surveys too sparse. As proxies for agricultural output, we use remote sensing indices developed by earth scientists to measure vegetation density, plant health, and metabolic activity. These vegetation indices have been shown to reliably predict crop yields across a range of settings (Running et al. 2004; Burke and Lobell 2017; Lobell et al. 2022; Asher and Novosad 2020). We use these indices to build a predictive model of crop yields, following Lobell et al. (2020), and calibrate it using aggregate statistics. In our data, each index individually predicts crop yields, and combining them forms better predictions than any individual index. From the combined model, we generate fitted values for each village in our sample.

We show three sets of results. First, we quantify the water pollution released by India's

"severely polluted" industrial sites, using the available monitoring station data. We show that there is a large, discontinuous increase in water pollution at these exact locations, raising omnibus measures of pollution in nearby rivers by 56 to 130 percent. The amount of water pollution released by these sites has not previously been estimated in publicly available sources.

Second, we find that our measure of crop yields, as predicted by remote sensing, is no lower in villages immediately downstream of high-polluting industrial sites than in comparable upstream villages in the same year. Confidence intervals for some specifications exclude yield reductions of 0.7 percent, suggesting that even the localized effects of industrial water pollution are small. Even in areas closest to rivers or near the largest industrial clusters, we do not detect pollution impacts.

Third, we find mixed evidence that farmers respond to industrial water pollution by switching irrigation sources from surface water to groundwater, and expanding irrigation overall. This suggests that the reason crop yields appear unharmed may be because farmers are adapting to industrial water pollution through costly input substitution.

Our study focuses on crop yields and does not imply that industrial water pollution is not costly. Even if output is unaffected, farmers may incur substantial averting expenditures in order to ensure that outcome. There are also many types of potential social costs that we do not quantify, including harm to ecosystems as well as to human health. Contaminated irrigation water may harm farmers and farm laborers who are exposed to it. Produce may take up heavy metals or other toxins, harming consumers even if yields are unaffected. These costs are outside the scope of this paper and important objects of future research.

This paper contributes evidence to three specific aspects of the costs of pollution. First, it studies the costs of water pollution from industrial sources. A large literature studies domestic water pollution in the context of drinking water (Olmstead 2010), while some papers study the effects of water pollution from all sources (Keiser and Shapiro 2019a) or agricultural sources (Brainerd and Menon 2014). Less evidence exists on industrial water pollution; exceptions include Ebenstein (2012) and Do et al. (2018), which find effects on cancer in China and infant mortality in India. Second, this paper studies how pollution affects the agricultural sector. Prior work on agriculture

focuses on the effects of air pollution (Burney and Ramanathan 2014; Aragón and Rud 2016), but there are physiological reasons to expect water pollution could harm crops as well. Third, this paper contributes to the effects of pollution specifically in low- and middle-income counties (Jayachandran 2009; Chen et al. 2013; Greenstone and Jack 2015; Adhvaryu et al. 2022).

This paper also contributes to a broader understanding of structural transformation and the relationship between industry and agriculture in low- and middle-income countries. Much existing literature focuses on input reallocation between sectors (Ghatak and Mookherjee 2014; Bustos et al. 2016), while this paper studies a non-pecuniary externality from industry to agriculture.

Finally, this paper makes progress in spatial computation methods for studying water pollution. In the United States, researchers can rely on the National Hydrography Dataset (Keiser and Shapiro 2019a; Keiser 2019; Andarge 2020; Taylor and Druckenmiller 2022; Jerch 2022; Flynn and Marcus 2021), the product of a vast modeling effort by the U.S. Geological Survey. Elsewhere, it can be difficult even to conceptually define upstream and downstream relationships, let alone compute them. We describe three specific challenges and how we overcome them. Alongside that of Garg et al. (2018), our approach may be useful to researchers studying water pollution in other settings.

2 Background on Water Pollution and Crop Growth

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).

By how much should we expect crop yields to fall downstream of the polluted industrial clusters? The answer will vary depends on the dose, exposure, and the particular mix of pollutants. We can provide a few reference points from agronomic studies on exposure to heavy metals. (Yang et al. 2021) found that a high dose of cadmium reduced total plant biomass of a Chinese medicinal plant by 50% within a year, relative to the control group that was not exposed. (Garzón et al. 2011) found that aluminium exposure reduced maize root growth by 40% within 24 hours of exposure. (Sharma and Sharma 1993) document chromium exposure reduced number of leaves in each wheat plant by 50%, while (Wallace et al. 1976) find that dry leaf yield in Bush bean plant decreased by 45% after chromium exposure.

A few small case studies suggest that the findings of 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 wastewater, 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.

How does water pollution reach crops? Possible exposure channels are through (a) surface water irrigation, using water pumped directly from a river; (b) surface water irrigation, using water from a canal that diverts water from the river; (c) groundwater irrigation, using water pumped from underground aquifers that may have been contaminated either through direct seepage or from surface water sources; or (d) soil contamination, from groundwater in areas with high water tables. Each of these exposure channels may produce different spatial patterns of treatment intensity, depending on topography, geology, soils, infrastructure, and irrigation practices, and they can operate over long time spans.

We cannot directly observe these exposure channels, since water and soil pollution is poorly monitored in India as in most of the world. Instead, we remain agnostic. Our research design captures the average effect of being downstream of a heavily-polluting industrial site, regardless of how the pollution arrives. The design is based on hydrological modeling of surface water flows, but surface water and groundwater are typically interconnected, and their flow gradients usually move together.

3 Research Design

Point sources of water pollution, such as industrial clusters, present a natural setting for a regression discontinuity design. Since water flows in only one direction, pollution levels immediately downstream of the point source will be discontinuously higher than pollution levels immediately upstream of the source.

Figure 1 illustrates this sharp discontinuity. It is an aerial photograph of one site in our 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. Although color is neither a sufficient nor necessary condition for any specific pollutant, the color difference confirms the presence of water from a different source, and color is correlated with water pollution. Remote sensing measures, which include visible light as well as a broader range of wavelengths, are becoming increasingly common in water quality monitoring (Gholizadeh et al. 2016).

3.1 Sample selection and treatment definition

The intuition for our research design is to compare agricultural outcomes in villages downstream of heavily-polluting industrial sites with those in villages upstream of the same sites. Although the idea is simple, translating it to precise definitions of "upstream" and "downstream" is less straightforward.

Our solution is illustrated in Figure 3. This figure shows our research design for one site in our sample: Bhillai-Durg, a major industrial city in the state of Chhattisgarh. The center of this industrial site is represented by the orange dot.

To construct a sample of villages for the RD design, we use hydrological modeling to find a "reference" flow line (shown in blue). This is a continuous streamflow path (from source to ocean) that satisfies three criteria: (1) it receives drainage from the industrial site, (2) it extends upstream into areas unaffected by the site, and (3) the point at which the drainage enters the stream is relatively close to the site itself. We construct this path by tracing the industrial site's drainage 25 km downstream and then following rivers both upstream and downstream of that point. Our sample is then formed by all villages within 20 km of the reference flow line. This radius gives us plenty of data to work with while focusing analysis on areas most likely to be affected by pollution.

To define the treatment status of villages, we compare their flow lengths with that of the industrial site, again calculated using hydrological modeling. The RD running variable – distance downstream of the industrial site – is the difference between these flow lengths. Villages are classified as downstream if they have a shorter flow length than the industrial site, and upstream otherwise.[1]

[1] We considered defining treatment status using elevation as the running variable, classifying villages with lower

Our approach captures the essential intuition of comparing "upstream" and "downstream" villages while solving three challenges. One challenge is that forgoing hydrological modeling can introduce severe measurement error. In settings where standardized hydrological data products are unavailable, researchers often simply snap the pollution sources to the nearest major river from a published shapefile (e.g., He et al. (2020)). But this method assumes all pollution impacts occur along a major river, which can miss the areas of greatest exposure for sources not located near a major river. It also inaccurately represents where pollution enters the river, resulting in false downstream and upstream classifications. Drainage does not flow orthogonally into the nearest river; it may enter the river somewhere far downstream, or it may not enter the *nearest* river at all. In our sample, we found that one industrial site drains to the Bay of Bengal, but its nearest major river in one shapefile flows in the opposite direction and drains to the Arabian Sea.

A second challenge is that there is no natural way to define an "upstream" set of villages without reference to a particular river or streamflow line. Upstream villages cannot be defined relative to a point source itself, since little land area drains directly into any given point outside of a river. To ensure a sample large enough for analysis, upstream villages must instead be defined relative to a point on a nearby flow line. This flow line must be close enough to maintain accurate links between source and exposure, but also major enough to yield a substantial upstream sample. We found that a flow line defined by a point 25 km downstream of the pollution source results in samples that satisfy both criteria.

The third challenge is that if downstream and upstream samples are selected in asymmetric ways, they may not be good counterfactuals for each other. We select downstream villages through the same process as the upstream villages – i.e., relative to the reference flow line, rather than the industrial site itself – creating a unified research design that avoids introducing mechanical discontinuities.

elevation than the industrial site as downstream (Asher et al. 2022). We found that this method produced small and insignificant RD estimates of the industrial sites on pollution concentrations. This weak "first stage" suggests that pollution exposure is better captured by flow length than elevation.

3.2 Regression discontinuity

Our main analyses estimate the causal effects of being immediately downstream of a heavily-polluting industrial site. We estimate standard RD regressions of the following form:

$$y_{ist} = \beta Downstream_{is} + \gamma Distance_{is} + \delta Distance_{is} \times Downstream_{is} + \alpha_{st} + \varepsilon_{ist} \quad (1)$$

in a sample consisting of the stacked upstream and downstream villages i corresponding to each industrial site s, across all observed years t.

The coefficient of interest is β, the local effect of being downstream of an industrial site. The running variable is downstream distance along the river flow path, defined such that each industrial site is at zero. Positive values indicate that a village is downstream of the industrial site; negative values indicate that the village is upstream. We 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 observations for the same industrial site in the same year. For pollution outcomes, all details are identical, except that i represents a water quality monitoring station instead of a village.

We estimate local linear regressions on each side of the cutoff without higher order polynomials, following Gelman and Imbens (2014). We report results using a range of bandwidths with a minimum value of 25 km. Smaller bandwidths might fail to include villages fully exposed to pollution, due to the way we construct our sample. We use a triangular kernel, which is optimal for estimating local linear regressions at a boundary (Fan and Gijbels 1996). We cluster standard errors by village to account for correlation across time. Clustering also accounts for repeated observations, when the same village appears more than once in the stacked sample for different industrial sites. Finally, we weight village observations by crop area so that our results represent the treatment effects for the average acre of cropland, which is more easily interpretable than effects for the average village.

The identifying assumption for this RD design is that the upstream patterns in pollution and agricultural outcomes would have continued smoothly downstream if the industrial site did not exist. Our samples represent continuous swaths of land area, making it *a priori* unlikely that there would

be discontinuities in either river pollution or agricultural outcomes. One way the assumption would be violated is if industrial sites were strategically located downstream of the best agricultural land. Most of the sites in our sample are part of cities and towns that arose through usual agglomeration processes, and we can test for discontinuities in land quality. Another way the assumption would be violated is if there is sorting of agricultural inputs or farmers themselves. Migration and/or disinvestment in downstream areas is possible, and we can test for it. These resources are unlikely to shift to the areas immediately upstream, rather than urban areas elsewhere, given India's rigid land and labor markets (Hsieh and Klenow 2009; Duranton et al. 2015).

3.3 Limitations of temporal variation

Our research design relies exclusively on cross-sectional variation because the variation we want to capture is predominantly spatial, not temporal. The timespan of pollution transport is unknown, and we want to capture the effects of pollution exposure through all possible channels. For example, diffusion through groundwater and accumulation in the soil can take years, decades, or more. Using temporal variation (e.g. with village or monitoring station fixed effects) would rule out these channels of transport that take longer to operate. Instead, we estimate the long-term cumulative effects of location relative to highly polluting industrial plants.

Temporal variation is also impractical in this setting because of low statistical power and high measurement error. The starkest variation in our context is spatial, not temporal – our causal identification is based on the location of industrial sites, which are extremely persistent and have not changed for decades. Most of these sites have grown over time, but this growth is correlated across sites over time as India has industrialized, leaving little useful variation. Available measures of industrial plant growth are noisy. The Economic Census gives the number of, and employment in, high-polluting plants in a town or village, but these variables are poor proxies for pollution and are known to suffer from data quality limitations (Bardhan 2013).

3.4 Impulse response functions

For some outcomes, we also use spatial impulse response functions to estimate non-local effects under stronger assumptions. The RD design estimates a local average treatment effect (LATE), which can tell us whether industrial pollution harms agriculture, and how large this harm is immediately downstream of industrial sites. However, it would be inappropriate to extrapolate RD estimates to all villages further downstream of industrial sites, because pollution tends to dissipate as the river flows downstream. Pollutants can break down, deposit on streambeds, or become diluted as a river collects runoff and joins other tributaries. To estimate the full effects of industrial sites over the course of a river, we use models of the following form:

$$y_{ist} = \gamma Distance_{is} + f(Distance_{is} \times Downstream_{is}) + \alpha_{st} + \varepsilon_{ist} \tag{2}$$

This equation is similar to an event study or distributed lag model, but in river space instead of time. The first term, $Distance_{is}$, controls for the linear trend of the outcome upstream of the industrial site. We then estimate a nonparametric function of distance on the downstream side. This function tells us the difference between the observed outcomes and the upstream trend, had it continued downstream.

To estimate this semiparametric model, we use a multistep process. First, we partial out site-by-year fixed effects α_{st} and obtain residuals. Second, we adjust for the upstream trend by regressing the residuals on $Distance_{is}$ for upstream observations only, obtaining fitted values for the downstream observations, and subtracting them from observed values. Third, we fit piecewise cubic splines to these adjusted values. We obtain 95% confidence intervals via cluster bootstrap, resampling districts with replacement and repeating the process for 1,000 iterations.

The assumption required for the spatial response function is considerably stronger than for the RD design. This design requires that the upstream trend can be extrapolated – that without the industrial sites, outcomes would have continued to follow the upstream trend downstream for as far as we estimate the function. This assumption is most likely to hold nearest to the downstream

cutoff, so the function is less reliable the further downstream we go. Despite these limitations, this design is the best available method to estimate the effects of industrial clusters away from the cutoff.

4 Predicting Yields Using Satellite Data

Our RD design requires agricultural outcome data at a high spatial resolution, at the level of fields or at least villages, across a large geographical area. The Indian government reports yearly agricultural data only at the administrative unit of districts, which span thousands of square kilometers. Survey and census microdata is rarely available in India or anywhere else and typically either lacks high-resolution spatial identifiers or is available for only a limited geographic extent.

Instead, we derive measures of crop yields from satellite data. Remote sensing data is now widely used in the scientific literature to measure crop yields (Running et al. 2004; Lobell et al. 2022), and it has started to be used in economics as well (Asher and Novosad 2020; Lobell et al. 2020). Satellite measures are known to predict yields well at small and large spatial scales, for many different crops, and in both high-income country settings (Hochheim and Barber 1998) and smallholder settings (Burke and Lobell 2017). In fact, Lobell et al. (2020) show that satellite measures can outperform farmer reporting and do at least as well as sub-plot crop cuts, as measured against the gold-standard measure of full-plot crop cuts.

The remote sensing literature has proposed a number of measures to proxy for crop yields. Rather than choose from among them, we follow Lobell et al. (2020) and put all available measures into a simple regression model. We then fit this model to the available district-level panel data on crop yields and generate predicted values for each village and year in our sample.

4.1 Vegetation indices

We use six vegetation indices (VIs). Five are used by Lobell et al. (2020): Normalized Difference Vegetation Index (NDVI), Green Chlorophyll Vegetation Index (GCVI), MERIS Terrestrial

Chlorophyll Index, Red-Edge $NDVI_{705}$ (NDVI705), and Red-Edge $NDVI_{740}$ (NDVI740). We also use the Enhanced Vegetation Index (EVI), following Asher and Novosad (2020) and Asher et al. (2022). NDVI and EVI are the two indices most commonly used in the scientific literature to proxy for agricultural output.

All VIs aim to capture the amount of photosynthetic activity in plants, which correlates with yields. Chlorophyll, the pigment that gives leaves their green color, absorbs much of the red light in the visible spectrum in healthy plants. Other cell structures of the plant reflect most of the near-infrared light in the invisible part of the electromagnetic spectrum. A healthy plant with high photosynthetic activity due to high amounts of chlorophyll will reflect less red light and more near-infrared light. Like cameras, satellite instruments capture the amount of light reflected in these different bands of the electromagnetic spectrum. Each VI is a function of different bands. NDVI uses red and near-infrared light; EVI is similar but uses additional information from the blue part of the electromagnetic spectrum to reduce atmospheric interference and the influence of background vegetation (Son et al. 2014). The other four VIs are variations on the same idea; each has shown useful in different settings (Burke and Lobell 2017).

4.2 Data

Satellite data. We extract minimum and maximum values of each VI during agricultural years 2015-17 from the Sentinel-2 MSI satellite[2] and aggregate them to village. Sentinel-2 is a satellite launched by the European Space Agency that records images at each point on Earth's land surface approximately once every 10 days, in a spatial resolution of 10 to 60 meters depending on band. The other major source of publicly available satellite imagery, NASA's Landsat 7, does not measure wavelengths in the ranges required to calculate NDVI705 and NDVI740. Maximum values of VIs are often found to be most strongly predictive of crop yields; minimum values (which likely occur during the off-season) may help control for background land cover factors (Asher and Novosad

[2]Accessed using Google Earth Engine, https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_HARMONIZED.

2020). India's agricultural year spans July 1 of the reference year through June 30 of the following year. We use years 2015-17 to correspond to the available district-level data.

To perform this calculation, we follow Lobell et al. (2020) as closely as possible. We read in each Sentinel-2 image taken of India between 1 July 2015 and 30 June 2018 and apply the quality assurance mask to remove clouds suggested by Google Earth Engine. To reduce noise, we also apply an agricultural land use mask from the Copernicus Global Land Service (CGLS) to ensure that only pixels of cropland are included in the sample. At each pixel, we calculate each VI at a 20m resolution for each image, then we find the minimum and maximum values of each VI during each agricultural year. Finally, we take means of the minimum and maximum values across all pixels with each village, to match with covariate data and improve computational tractability. For district-level VIs, we take means of village-level values, weighting villages by agricultural land area from the population census.

District-level agricultural outcomes. We calculate crop yields weighted by price, which we refer to as the "revenue value of yield," from the District Level Database compiled by ICRISAT.[3] This data contains information on crop area planted, output and prices for 16 major crops, for 571 districts across 20 states from 1990-2017. Price data covers about 79% of all area under cultivation. Revenue value of yield is calculated by multiplying the quantity of each crop by the (time-invariant) mean price for that crop in that district between 1990-2002. For districts without price data, we impute the state mean if available or the national mean otherwise.

4.3 Predictive model

We first verify that our calculated VIs are individually predictive of, and positively correlated with, crop yields. To do so, we regress log revenue value of yield on the log of the difference between maximum and minimum values of each VI, following Asher and Novosad (2020). This is a district-level cross-sectional regression; we omit spatial fixed effects since our final research design relies

[3] http://data.icrisat.org/dld/src/crops.html

on spatial variation. Results are shown in the first four columns of Table 1. NDVI, EVI, GCVI, and MTCI are each positively correlated with log revenue and individually explain a substantial fraction (between 6 and 21 percent) of the variation in log revenue.

Next, we fit our predictive model: We regress log revenue value of yield on all six VIs, with maximum and minimum values entering separately and linearly, following Lobell et al. (2020). Results are shown in column (5) of Table 1. Individual coefficients lack an intuitive interpretation, since each is conditional on all the others. However, the explanatory power of this regression far exceeds any of the individual VIs, with an R^2 of 0.39. For comparison, Lobell et al. (2020) report an R^2 of 0.58 in plot-level data of homogeneous crops in a small geographical region. Considering that our data is spread across a much larger region with heterogeneous crops, our model performance appears good.

Finally, we obtain village-level predicted yields by fitting this estimated model to our village-level VI data, for all years of our sample. One limitation of this approach is that we cannot measure the performance of our district-level model in village-level data, since no systematic village-level data on agricultural outcomes is available. In addition, the relationship between reflectance and yields varies by crop, so our model would surely be improved by controlling for crop shares. Crop identification maps exist for the United States (i.e., the USDA's Cropland Data Layer) and are under development for India, but none are publicly available yet. Census data on village amenities lists the major crops in each village, but data quality is too low to be useful.

5 Other Data and Summary Statistics

5.1 Data Sources

5.1.1 Industrial sites

India's 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 Envi-

ronmental Assessment of Industrial Clusters" (Central Pollution Control Board 2009). We identified the geolocation of each site using Google Maps and other publicly available reference information. These sites are displayed as orange dots in Figure 2.

The CPCB document also contains numerical scores for air, water, and land pollution, and an overall score, each out of 100. Land pollution refers to toxic waste, which can also contaminate groundwater. Details of the scoring methodology are provided in a companion document (Central Pollution Control Board 2009). 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). Our sample consists of 48 such sites that had a "severe" rating in land or water pollution in 2009 and for which our sample selection procedure yielded at least one upstream and downstream village per site.

5.1.2 Pollution measurements

We use data of water pollution measurements along rivers in India collected by the CPCB. The initial dataset, collected and published by Greenstone and Hanna (2014), includes monthly observations from 459 monitoring stations along 145 rivers between the years 1986 and 2005. We extend this data by downloading yearly pollution readings for the same stations from 2006-2012 from the CPCB website. We construct yearly averages for the pre-2005 data and append these to the newly downloaded data.

This raw dataset includes a noisy location measure as well as river name and a description of the sampling location. We verified, refined, or corrected the geolocation of each station by manually cross-referencing these contextual variables with Google Maps, CPCB documents, and other publicly available reference information. The locations of these stations are displayed as green dots in Figure 2.

Many water quality parameters have been collected by the CPCB at some point. However, only a few parameters are measured consistently. We focus on three common omnibus measures that proxy for a wide range of pollutants: chemical oxygen demand (COD), biochemical oxygen

demand (BOD), and dissolved oxygen saturation (DO). COD is a standardized laboratory test that serves as an omnibus measure of organic compounds, which industrial plants typically generate in high quantities. BOD is a related but narrower test. COD and BOD are the Indian government's top priority in regulating industrial wastewater (Duflo et al. 2013), while DO is widely used in research (Keiser and Shapiro 2019a). We also show results for a number of less consistently reported measures.

5.1.3 Village covariates and boundaries

For baseline village covariates, we use the Population Census of 2001, which includes more than 200 variables on population, employment, amenities, and infrastructure. We obtain cleaned Census data along with geospatial data on village boundaries from NASA's Socioeconomic Data and Applications Center.[4] Because villages and towns sometimes split or merge, we use consistent definitions from the Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG) provided by the Development Data Lab.[5] The SHRUG provides an identifier called a "shrid" for a group of contiguous villages or towns that can be combined into unchanged spatial entities over several decades. Almost 96% of villages from the 2001 population match a single shrid and do not require spatial aggregation. For the rest, we dissolve polygons boundaries to obtain shrid boundaries, and aggregate administrative data over the villages within each shrid.

5.2 Hydrological modeling

We use the following procedure to match villages and pollution monitoring stations to industrial sites and assign river distances and treatment status.

Flow length raster. We 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

[4] Available at https://sedac.ciesin.columbia.edu/data/set/india-india-village-level-geospatial-socio-econ-1991-2001

[5] Available at https://www.devdatalab.org/shrug_download/

this DEM, we use the Spatial Analyst tools in ArcGIS Pro to fill sinks, create a flow direction raster (using the D8 method), and derive a flow length raster. This raster gives the distance along rivers that a particle released at each cell must travel to reach the ocean (or the edge of the raster).

Sample selection. To define the sample of villages for each industrial site, we first create a reference flow line. We use the Trace Downstream tool in ArcGIS Pro to find the site's flow path, i.e., the route that effluent released at the site must follow to reach the ocean. We then find the point on this flow path that is 25 km downstream of the site (the upper yellow dot in Figure 3). Next, we use the Watershed tool in ArcGIS Pro to find the area that drains into that point. We find the flow lengths of all villages within this watershed by intersecting the watershed polygon with village centroids and matching village centroids to the flow length raster. We identify the longest possible flow path within this watershed by choosing the village at the 95th percentile of flow length within this set (the lower yellow dot). We use the 95th percentile instead of the maximum to avoid erroneous values that sometimes arise at the edges of watershed polygons. Finally, to define the sample, we find the flow path of the chosen "headwater" village, generate a 20-km buffer around each flow path, and intersect this buffer with village centroids.

Village distance and treatment status. To calculate distances for the RD design, we project village centroids and monitoring stations into one-dimensional river space, snapping them to the nearest point along the reference flow line. We then find the flow length (i.e., to the ocean) of each snapped point by matching it to the flow length raster. We construct distance, the running variable, as the difference in flow lengths between each village or monitoring station and its corresponding industrial site. We also construct a downstream indicator variable equaling one if the distance variable is positive, meaning that the village or station is downstream of the industrial site. We also calculate the perpendicular distance from the original village centroid to the flow line, as a control and for heterogeneity analysis.

5.3 Continuity tests and summary statistics

We provide summary statistics in Table 2 for our main outcome variables on pollution and agricultural output, and in the first column of Appendix Table 6 for covariates.

To assess the credibility of our research design, we test a range of covariates for continuity at the threshold of being downstream of the industrial site. If the identification assumption is true, we should not see discontinuous jumps in the values of other village characteristics that are fixed or unlikely to be affected by pollution. We test for continuity by running RD regressions in the form of Equation 1 with each covariate on the left-hand side. For the RD design, covariate means do not need to be equal upstream and downstream; they only need to vary continuously as the river passes the industrial site.

We group covariates into several categories: (1) physical characteristics, (2) potential yields estimated for common crops, (3) commercial and public amenities, and (4) demographic characteristics. Physical characteristics and potential yields are time-invariant and cannot be affected by water pollution, so they are the "purest" tests. In contrast, amenities and demographics could potentially respond to water pollution if the economic impacts are large enough. For these variables, a discontinuity could represent a genuine outcome rather than evidence of pre-existing difference. Still, we include them because they are important characteristics of villages and we expect any endogenous response to be small compared with overall patterns.

Figure 4 shows visual evidence of continuity for a selection of these covariates. For context, we first plot a histogram of village observations. The usual density test of McCrary (2008) is unnecessary since our sample is based on land area, which by definition has a continuous density in space; villages cannot manipulate their locations relative to the cutoff. Other plots in Figure 4 suggest that elevation, potential yields (standardized and averaged across crops), distance to nearest canal, village population, and share of population in scheduled castes and scheduled tribes are all roughly continuous.

Standard errors and RD point estimates for these covariates and many others are shown in Appendix Table 6 using a range of bandwidths. Across the 31 variables and 3 bandwidths we test, 12

estimates are statistically significant at a 10% level, in line with expectations. Taken together, there is little evidence to suggest that agricultural outcomes would be different immediately downstream of the industrial sites if they did not exist. It also does not appear that commercial and public amenities or demographic characteristics of villages are affected by being downstream of these industrial sites. In robustness checks, we control for all these covariates.

6 Results

6.1 Pollution

We 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.

Figure 5 visualizes our main results for pollution. The left side shows regression discontinuity plots for three water quality parameters that are both widely reported and associated with industrial pollution: chemical oxygen demand (COD), biological oxygen demand (BOD), and dissolved oxygen (DO). The graphs plot mean values of each parameter within quantile bins of distance from the industrial site; each dot represents approximately 260 observations. Positive distance values indicate that the monitoring station is downstream of the industrial site, and negative values are upstream stations. Before binning, values are log-transformed and adjusted for site-by-year fixed effects. We also fit cubic splines to show overall patterns.

All three parameters show a discontinuous increase in pollution at the exact location of the industrial sites. COD and BOD increase; these parameters are undesirable, with higher levels indicating worse water quality. The decrease in DO also indicates an increase in pollution; this parameter is desirable, with lower levels indicating worse water quality.

Graphs on the right side of Figure 5 show that water pollution dissipates as the river flows downstream. These graphs plot spatial impulse response functions for each parameter, showing how industrial clusters affect river pollution over the course of the river. For all three parameters, the increase in pollution is greatest immediately after the industrial site. It then steadily falls and

rejoins the trend implied by the upstream curve no more than 100 km from the industrial site.

Table 3 quantifies these results. It reports RD estimates from Equation 1 estimated separately for each parameter, for bandwidths of 25, 50, and 100 km. Dependent variables are listed in rows; each cell shows the estimated coefficient on the Downstream indicator variable, controlling for distance on each side of the industrial site along with site-by-year fixed effects.

The estimates are quantitatively large. For example, the estimate of 80.6 for COD (with a 50-km bandwidth) implies that the average "severely polluted" industrial site nearly doubles pollution levels in nearby rivers. Confidence intervals exclude zero at a 95% level for all three parameters at bandwidths of 50 and 100 km. Estimates using a bandwidth of 25 km are less precise but have very similar point estimates; standard errors shrink as bandwidths increase and more data enters the sample.

Appendix Table 7 reports RD results for 16 additional water pollutants available in CPCB data. These pollutants are measured less frequently, so many of the estimates are imprecise. However, the evidence suggests that nearly every reported pollutant doubles or triples in concentration (or increases by around 1 standard deviation) immediately downstream of industrial sites. This is true for measures of salinity (electrical conductivity and presence of ions like calcium, chloride, magnesium, and sodium), nutrients (nitrates, nitrites, potassium, and sulphates), acidity (pH), and other omnibus measures (total solids and turbidity).

No data is available to directly measure heavy metals or toxic organic chemicals, which are likely the most concerning pollutants for crop growth. However, our research design is based around the industrial sites that are likely some of the greatest sources of these water pollutants in India if not the world, so it is reasonable to expect heavy metals and organic compounds to rise in tandem with other parameters at these locations. Most importantly, the fact that essentially every observed pollutant increases dramatically at the precise locations of these industrial sites represents a strong "first stage" that gives us confidence that our research design is indeed capturing the pollution exposure we want it to.

6.2 Agricultural outcomes

Having shown that industrial sites increase pollution, we investigate how this pollution affects agricultural production in downstream villages, using our measure of crop yields derived from satellite data.

Figure 6 visualizes our main result for crop yields. It shows an RD plot similar to those for pollution, but using the predicted log revenue value of yield, by village and year. The plot does not show a discontinuity at the industrial site. Despite increasing water pollution drastically, industrial sites do not seem to affect downstream crop yields.

Table 4 quantifies this result. As before, Panel A reports RD estimates for predicted crop yield for multiple bandwidths (in columns). The point estimate using a 50-km bandwidth is 0.009, implying that crop yields are 0.9 percent *higher* immediately downstream of a severely-polluting industrial site. This apparent increase is not statistically different from zero. The 95% confidence interval allows us to reject reductions in crop yields larger than 0.7 percent. Other bandwidths yield results that are less precise but still small in magnitude.

Panels B-D of Table 4 report results from variations on the main specification. Panel B controls for the distance from village to river flow line. Panel C controls for the full set of pre-treatment variables tested in Appendix Table 6. Panel D controls for irrigation-related agricultural input variables listed in Table 5. All these specifications produce similar results as the main specification. None of the estimates are statistically different from zero, and the point estimate with the largest magnitude is -0.024, a 2.4 percent reduction.

6.3 Heterogeneity

Might our research design still examine too broad of an area? To zero in on the precise areas likely to have the greatest pollution exposure, we conduct heterogeneity analyses along three dimensions. First, 8 examines heterogeneity by distance to river (i.e., flow line). Villages closer to the affected river are more likely to be directly affected by pollution, either through groundwater or through river irrigation. Next, 8 examines heterogeneity by total employment in highly polluting industries

within the industrial site, as calculated from the Economic Census. Even among the severely polluting sites in our sample, those with greater concentration of employment in the most-polluting industries may generate more pollution. Finally, 10 examines heterogeneity by irrigation source (i.e., whether the village has any cropland irrigated by canals, wells, or rivers). Which irrigation sources deliver the most water pollution is unknown, but some may deliver more than others.

Across all subgroups, there are no detectable effects of industrial sites on crop yields. All magnitudes are small, and all but one are statistically insignificant. There are not even any suggestive patterns across point estimates.

6.4 Agricultural inputs

We next look at whether farmers adjust irrigation and other agricultural inputs in response to industrial water pollution. Effects on agricultural inputs can provide a fuller description of the potential costs of pollution. Even though crop yields are not harmed much, that may be a net result of costly adaptation choices, as farmers reallocate factors of production toward or within agriculture in order to maintain crop yields.

Table 5, Panel B reports RD estimates for a set of agricultural inputs. Labor, as measured by the share of employment in agriculture, does not change much immediately downstream of heavily-polluting industrial sites (for one bandwidth the point estimate is statistically significant but still small). Neither does land, as measured by crop area under cultivation (per capita).

However, irrigation inputs do appear to respond to industrial water pollution, at least for some bandwidths. The share of crop area under irrigation increases up to 3 percentage points, the number of villages that irrigate from rivers and canals fall by up to 7 and 11 percentage points, and the number of villages that irrigate from wells increases by up to 4 percentage points. This evidence is more mixed because the estimates are not consistent across bandwidths. But overall, it appears that farmers may respond to industrial water pollution by (a) switching irrigation sources away from surface water and toward groundwater, and (b) expanding irrigation overall.

If this is true, it suggests two things: First, surface water is the more likely channel of exposure

through which pollution reaches farms. Second, the null effect on crop yields is masking larger welfare costs of pollution: Perhaps the only reason crop yields are unharmed by industrial water pollution in equilibrium as a result of costly input substitution, as farmers drill more wells and use more energy to pump groundwater.

7 Discussion

7.1 Contextualizing the results

Our results suggest that crop yields are not detectably harmed by industrial water pollution. It is still possible that certain villages near certain industrial sites experience damages. But on average, our most precise estimates can reject declines in crop yields of more than 0.7 percent.

How does this magnitude compare with other kinds of impacts to crop yields? Estimates are larger for many other shocks and interventions. Yields fall 4 percent in response to a one standard deviation increase in average temperature (Colmer 2021), 2 to 8 percent in response to heat waves (Heinicke et al. 2022), 3 to 10 percent in response to a 20-day delay in monsoon arrival (Amale et al. n.d.), and 20 to 36 percent in response to air pollution (Burney and Ramanathan 2014). Productivity gains from crop germplasm improvement in the Green Revolution are estimated at 0.5 to 1.0 percent *per year* over multiple decades (Pingali 2012).

In addition, our estimates likely represent an upper bound on the overall impacts of industrial water pollution on crops, for two reasons. First, our study focuses on the most highly polluting industrial sites in India, so the effects of other pollution sources should be smaller. Second, our RD regressions estimate a *local* treatment effect immediately downstream of heavily-polluting industrial sites. Since pollution dissipates further away from the sites, the effects further downstream will be smaller.

Even if we take the lower end of the least-precise confidence interval in Table 4, the largest pollution effect that could possibly be consistent with our estimates is a 9 percent fall in crop yields. Damages of this magnitude would indeed be harmful for farmers in the affected area. But this upper

bound would apply only to a very small region. Assuming crop yield impacts approximately scale with pollution concentrations, crops more than 50 to 100 km downstream of the clusters would be essentially unaffected.

7.2 Explaining the results

It may be puzzling – and at odds with the agronomy literature – that near some of the largest point sources of industrial water pollution in the world, crops seem not to be harmed. We propose and discuss six hypotheses to explain our results.

Hypothesis 1: Farmers adjust agricultural inputs to avert pollution damage. This is the explanation for which we find mixed evidence. In some specifications, farmers downstream of industrial sites appear to irrigate more of their crops and shift from surface water to groundwater sources. Like households that adopt air conditioning to avoid damage from heat, farmers may substitute inputs to avoid pollution damages they would otherwise suffer. In this case, the welfare cost of pollution would be found not in the dose-response effects but rather in these averting expenditures.

Hypothesis 2: These specific yield impacts are not well suited to detection by remote sensing. Many papers in the economics and scientific literatures have found satellite-derived measures to be useful proxies for crop yields and agricultural output, including for answering causal questions. For example, Asher et al. (2022) find a positive effect of canal construction on EVI in India. However, many questions and uncertainties remain about their capabilities. One possibility is that vegetation indices are simply not well-suited to pick up the specific effects of industrial water pollution on crops. This seems unlikely, since many of the agronomy studies on water pollution specifically report negative impacts to leaf size and color, characteristics that vegetation indices are well-tailored to measure. Another possibility is that farmers adjust crop choice in response to pollution exposure. Vegetation indices are affected by vegetation type in addition to crop health, so if farmers switch to new crops with greater baseline biomass or leaf canopy, it could offset the direct harms from pollution. Controlling for crop type could rule out this concern, but no crop

classification datasets are yet publicly available.

Appendix Table 11 shows the results of our main analysis applied to the best available data that directly observes crop yields, the ICRISAT district-level data. As expected, estimates are too noisy and imprecise to be useful.

Hypothesis 3: Pollution harms output quality rather than quantity. It is possible that industrial water pollution does harm crops, but only in ways that affect crop quality rather than quantity. For example, a crop such as rice might absorb heavy metals, bringing adverse health effects to consumers but leaving yield unaffected. The welfare consequences of quality effects are harder to measure. Obvious quality effects such as discoloration may capitalize into prices, but other quality effects may not. Studying effects of pollution on crop revenues rather than yields would address this issue; we again attempt to do this in Appendix Table 11 but find results to be highly imprecise.

Hypothesis 4: Industrial water pollution has beneficial components that balance the harms. Industrial effluent often includes salinity, heavy metals, and other components that are known to harm crops. However, they can also include nitrates, phosphates, and potassium, which can benefit plants as nutrients. It is possible that the net effects of industrial effluent are near zero, even if individual components have positive and negative effects.

Hypothesis 5: Farm-level doses are lower than observed pollution levels. Perhaps the high levels of industrial pollution measured in rivers are not as large at the point when actually applied to crops. The most direct channels of pollution transport are rivers and canals, but these surface water sources irrigate a relatively small share of land. Most irrigation water in India is pumped from wells, and the transport and fate of pollutants in groundwater is complex. Perhaps industrial effluent filters through enough layers of soil and rock that pollutants are removed, remediated, or diluted before being taken up by crops.

Hypothesis 6: Case studies exhibit publication bias. Although a number of studies in agronomy have shown significant impacts of industrial water pollution on crops, it is possible that the studies available in the published literature are unrepresentative of the true overall effects of pollution. This could happen in two ways. One is site selection: Perhaps the cases researchers choose to

study are extreme outliers in pollution concentrations, directness of crop exposure, or vulnerability of specific crops to specific pollutants, and impacts to crops more generally are smaller. The other way is file drawer bias: Even if true pollution effects are small, sampling variation will produce larger results for some studies, and other studies are abandoned without publication.

8 Conclusion

This paper studies the effects of industrial water pollution on agriculture. We examine 48 industrial sites in India identified by the government as "severely polluting" and estimate the costs of their pollution to downstream agriculture. Our regression discontinuity research design exploits the unidirectional flow of water pollution along with the location of these severely polluted industrial sites. To overcome the limitations placed by spatially aggregated administrative data on agricultural output, we build predictive models of crop yields from vegetation indices in satellite data. Such models have been shown to perform well in predicting yields both in the scientific and economics literature, and we verify that they predict agricultural yields within our sample too. We also use hydrological modeling to model areas of pollution exposure and choose counterfactuals.

We describe three sets of results. First, the location of these industrial sites coincides with a large, discontinuous jump in water pollution in nearby rivers. Second, crop yields are no lower in villages immediately downstream of these sites than comparable villages immediately upstream of the same sites, in the same year. Third, we find some evidence that farmers adjust irrigation inputs to avoid pollution damages: in some specifications, downstream villages irrigate more overall, are less likely to use surface water irrigation, and are more likely to use groundwater irrigation.

We propose six hypotheses to explain findings. Besides adjusting inputs, it is possible that the specific types of yield impacts caused by industrial water pollution are not well-suited to detection via remote sensing, that pollution harms output quality rather than quantity, that industrial pollution has beneficial components for agriculture that balance the harms, that farm-level pollution levels are lower than river observations, or that the case studies in the agronomic literature exhibit publication

bias. Due to data limitations, we leave the resolution of these explanations to future research.

9 References

Adhvaryu, Achyuta, Namrata Kala, and Anant Nyshadham. 2022. "Management and Shocks to Worker Productivity." Journal of Political Economy 130 (1): 1–47.

Amale, Hardeep Singh, Pratap Singh Birthal, and Digvijay Singh Negi. n.d. "Delayed Monsoon, Irrigation and Crop Yields." Agricultural Economics n/a (n/a). Accessed December 24, 2022.

Andarge, Tihitina. 2020. "Effect of Incomplete Information on Ambient Pollution Levels." https://drive.google.com/file/d/1Vov2o3b-ACS3mgN7n_bMqgmFqVKFSqqT/view?usp=embed_facebook.

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

Arceo, Eva, Rema Hanna, and Paulina Oliva. 2016. "Does the Effect of Pollution on Infant Mortality Differ Between Developing and Developed Countries? Evidence from Mexico City." The Economic Journal 126 (591): 257–80.

Asher, Sam, Alison Campion, Douglas Gollin, and Paul Novosad. 2022. "The Long-Run Development Impacts of Agricultural Productivity Gains: Evidence from Irrigation Canals in India."

Asher, Sam, and Paul Novosad. 2020. "Rural Roads and Local Economic Development." American Economic Review 110 (3): 797–823.

Bardhan, Pranab. 2013. "The State of Indian Economic Statistics: Data Quantity and Quality Issues." Public {Lecture}. University of California, Berkeley.

Brainerd, Elizabeth, and Nidhiya Menon. 2014. "Seasonal Effects of Water Quality: The Hidden Costs of the Green Revolution to Infant and Child Health in India." Journal of Development Economics 107: 49–64.

Burke, Marshall, and David B. Lobell. 2017. "Satellite-Based Assessment of Yield Variation and Its Determinants in Smallholder African Systems." Proceedings of the National Academy of Sciences 114 (9): 2189–94.

Burney, Jennifer, and V. Ramanathan. 2014. "Recent Climate and Air Pollution Impacts on Indian Agriculture." Proceedings of the National Academy of Sciences 111 (46): 16319–24.

Bustos, Paula, Bruno Caprettini, and Jacopo Ponticelli. 2016. "Agricultural Productivity and Structural Transformation: Evidence from Brazil." American Economic Review 106 (6): 1320–65.

Central Pollution Control Board. 2009. "Criteria for Comprehensive Environmental Assessment of Industrial Clusters."

Chen, Yuyu, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. 2013. "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 110 (32): 12936–41.

Colmer, Jonathan. 2021. "Temperature, Labor Reallocation, and Industrial Production: Evidence from India." American Economic Journal: Applied Economics 13 (4): 101–24.

Do, Quy Toan, Shareen Joshi, and Samuel Stolper. 2018. "Can Environmental Policy Reduce Infant Mortality? Evidence from the Ganga Pollution Cases." Journal of Development Economics 133 (September 2016): 306–25.

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

Duranton, Gilles, Ejaz Ghani, Arti Grover Goswami, and William R. Kerr. 2015. "Effects of Land Misallocation on Capital Allocations in India." Policy {Research} {Working} {Paper};{No}. 7451. Washington, DC: World Bank.

Ebenstein, Avraham. 2012. "The Consequences of Industrialization: Evidence from Water Pollution and Digestive Cancers in China." The Review of Economics and Statistics 94 (1): 186–201.

Fan, Jianqing, and Irene Gijbels. 1996. "Local Polynomial Modelling and Its Applications." Monographs on Statistics and Applied Probability 66.

FAO. 2018. Water for Sustainable Food and Agriculture: A Report Produced for the G20 Presidency of Germany. Food & Agriculture Org.

Flynn, Patrick, and Michelle M. Marcus. 2021. "A Watershed Moment: The Clean Water Act and Infant Health." Working {Paper}. Working Paper Series. National Bureau of Economic

Research.

Garg, Teevrat, Stuart E. Hamilton, Jacob P. Hochard, Evan Plous Kresch, and John Talbot. 2018. "(Not so) Gently down the Stream: River Pollution and Health in Indonesia." Journal of Environmental Economics and Management 92 (November): 35–53.

Garzón, Teresa, Benet Gunsé, Ana Rodrigo Moreno, A. Deri Tomos, Juan Barceló, and Charlotte Poschenrieder. 2011. "Aluminium-Induced Alteration of Ion Homeostasis in Root Tip Vacuoles of Two Maize Varieties Differing in Al Tolerance." Plant Science 180 (5): 709–15.

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

Ghatak, Maitreesh, and Dilip Mookherjee. 2014. "Land Acquisition for Industrialization and Compensation of Displaced Farmers." Journal of Development Economics 110: 303–12.

Gholizadeh, Mohammad Haji, Assefa M. Melesse, and Lakshmi Reddi. 2016. "A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques." Sensors (Basel, Switzerland) 16 (8): 1298.

Greenstone, Michael, and Rema Hanna. 2014. "Environmental Regulations, Air and Water Pollution, and Infant Mortality in India." American Economic Review 104 (10): 3038–72.

Greenstone, Michael, and B. Kelsey Jack. 2015. "Envirodevonomics: A Research Agenda for an Emerging Field." Journal of Economic Literature 53 (1): 5–42.

He, Guojun, Shaoda Wang, and Bing Zhang. 2020. "Watering Down Environmental Regulation in China*." The Quarterly Journal of Economics 135 (4): 2135–85.

Heinicke, Stefanie, Katja Frieler, Jonas Jägermeyr, and Matthias Mengel. 2022. "Global Gridded Crop Models Underestimate Yield Responses to Droughts and Heatwaves." Environmental Research Letters 17 (4): 044026.

Hochheim, K. P., and D. G. Barber. 1998. "Spring Wheat Yield Estimation for Western Canada Using NOAA NDVI Data." Canadian Journal of Remote Sensing 24 (1): 17–27.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China

and India*." The Quarterly Journal of Economics 124 (4): 1403–48.

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

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

Jerch, Rhiannon L. 2022. "The Local Benefits of Federal Mandates: Evidence from the Clean Water Act."

Keiser, David. 2019. "The Missing Benefits of Clean Water and the Role of Mismeasured Pollution." Journal of the Association of Environmental and Resource Economists, July.

Keiser, David, and Joseph S Shapiro. 2019a. "Consequences of the Clean Water Act and the Demand for Water Quality*." The Quarterly Journal of Economics 134 (1): 349–96.

Keiser, David, and Joseph S. Shapiro. 2019b. "US Water Pollution Regulation over the Past Half Century: Burning Waters to Crystal Springs?" Journal of Economic Perspectives 33 (4): 51–75.

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

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

Lobell, David, George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic, and Siobhan Murray. 2020. "Eyes in the Sky, Boots on the Ground: Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis." American Journal of Agricultural Economics 102 (1): 202–19.

Lobell, David, Stefania Di Tommaso, and Jennifer A. Burney. 2022. "Globally Ubiquitous Negative Effects of Nitrogen Dioxide on Crop Growth." Science Advances 8 (22): eabm9909.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity

Design: A Density Test." Journal of Econometrics 142 (2): 698–714.

Mohan, Vishwa. 2021. "India's 88 Industrial Clusters Present a Bleak Picture of Air, Water and Land Contamination, Says CSE Report." The Times of India, February.

Möller-Gulland, Jennifer. 2018. "Toxic Water, Toxic Crops: India's Public Health Time Bomb." Circle of Blue.

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

Olmstead, Sheila M. 2010. "The Economics of Water Quality." Review of Environmental Economics and Policy 4 (1): 44–62.

Pingali, Prabhu L. 2012. "Green Revolution: Impacts, Limits, and the Path Ahead." Proceedings of the National Academy of Sciences 109 (31): 12302–8.

Reddy, V. Ratna, and Bhagirath Behera. 2006. "Impact of Water Pollution on Rural Communities: An Economic Analysis." Ecological Economics 58 (3): 520–37.

Running, Steven W., Ramakrishna R. Nemani, Faith Ann Heinsch, Maosheng Zhao, Matt Reeves, and Hirofumi Hashimoto. 2004. "A Continuous Satellite-Derived Measure of Global Terrestrial Primary Production." BioScience 54 (6): 547–60.

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

Sharma, D. C., and C. P. Sharma. 1993. "Chromium Uptake and Its Effects on Growth and Biological Yield of Wheat." Cereal Research Communications 21 (4): 317–22. https://www.jstor.org/stable/23783985.

Son, N. T., C. F. Chen, C. R. Chen, V. Q. Minh, and N. H. Trung. 2014. "A Comparative Analysis of Multitemporal MODIS EVI and NDVI Data for Large-Scale Rice Yield Estimation." Agricultural and Forest Meteorology 197 (October): 52–64.

Taylor, Charles A., and Hannah Druckenmiller. 2022. "Wetlands, Flooding, and the Clean Water Act." American Economic Review 112 (4): 1334–63.

Vyas, Ananya. 2022. "Explainer: What Is Causing the Mass Death of Fish in India's Water Bod-

ies?" Text. Scroll.in.

Wallace, A., S. M. Soufi, J. W. Cha, and E. M. Romney. 1976. "Some Effects of Chromium Toxicity on Bush Bean Plants Grown in Soil." Plant and Soil 44 (2): 471–73.

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

Yang, Jiyuan, Hui Sun, Jihong Qin, Xiaoqin Wang, and Wenqing Chen. 2021. "Impacts of Cd on Temporal Dynamics of Nutrient Distribution Pattern of Bletilla Striata, a Traditional Chinese Medicine Plant." Agriculture 11 (7): 594.

10 Figures

Figure 1: Satellite photo showing a discontinuity in river color at the outlet of the Nazafgarh Drain Basin on the Yamuna River, just north of New Delhi. (Source: Sentinel 2, taken on October 2, 2017.) ↩

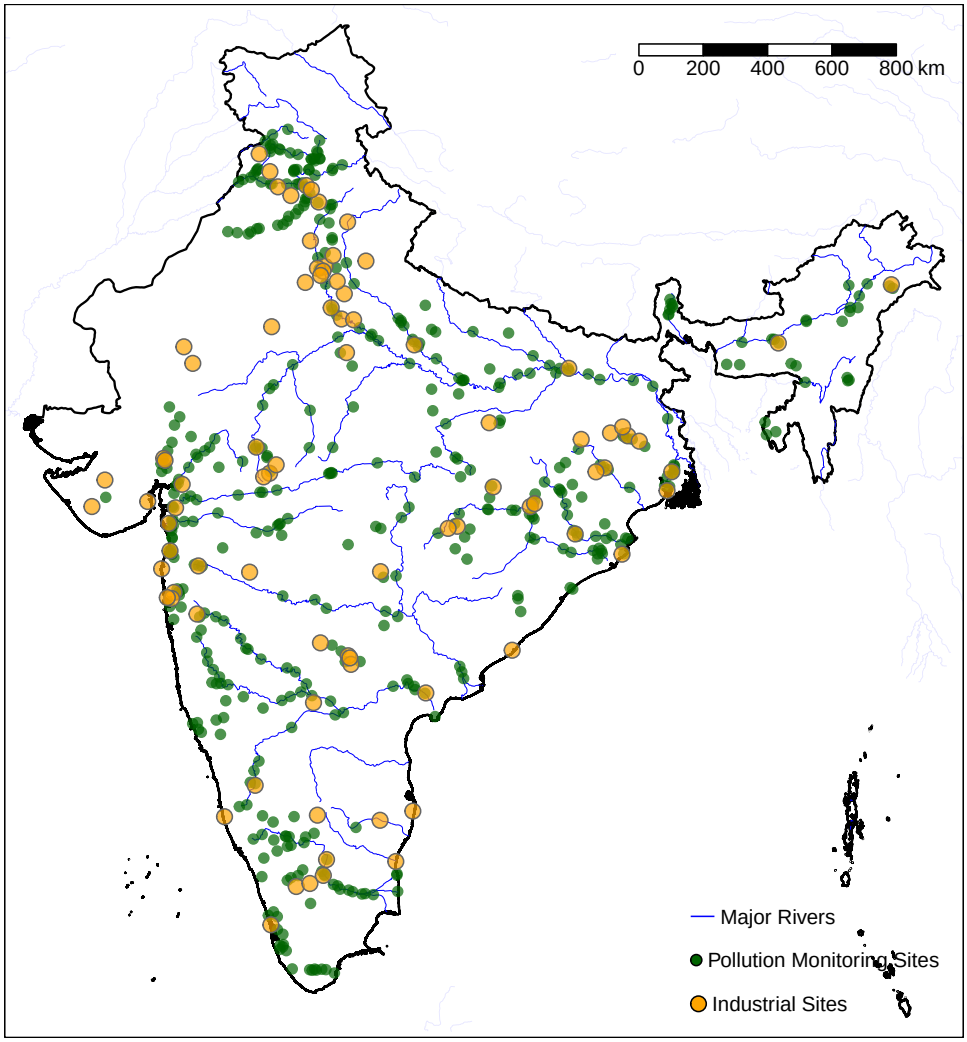

Figure 2: Locations of "severely polluted" industrial sites (orange dots) and water pollution measurement stations (green dots).↩

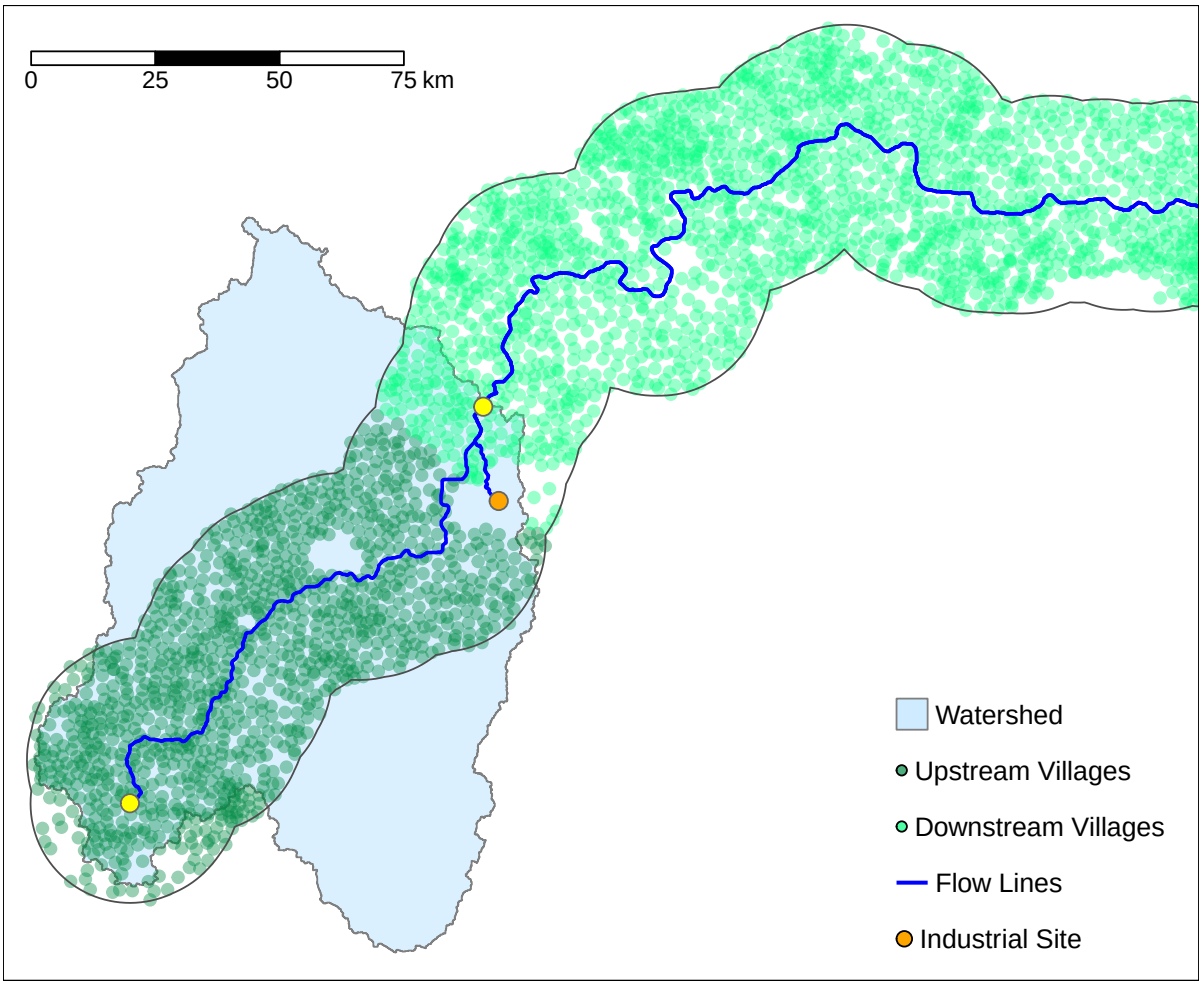

Figure 3: Illustration of the sample selection and treatment assignment for our main research design. ↩

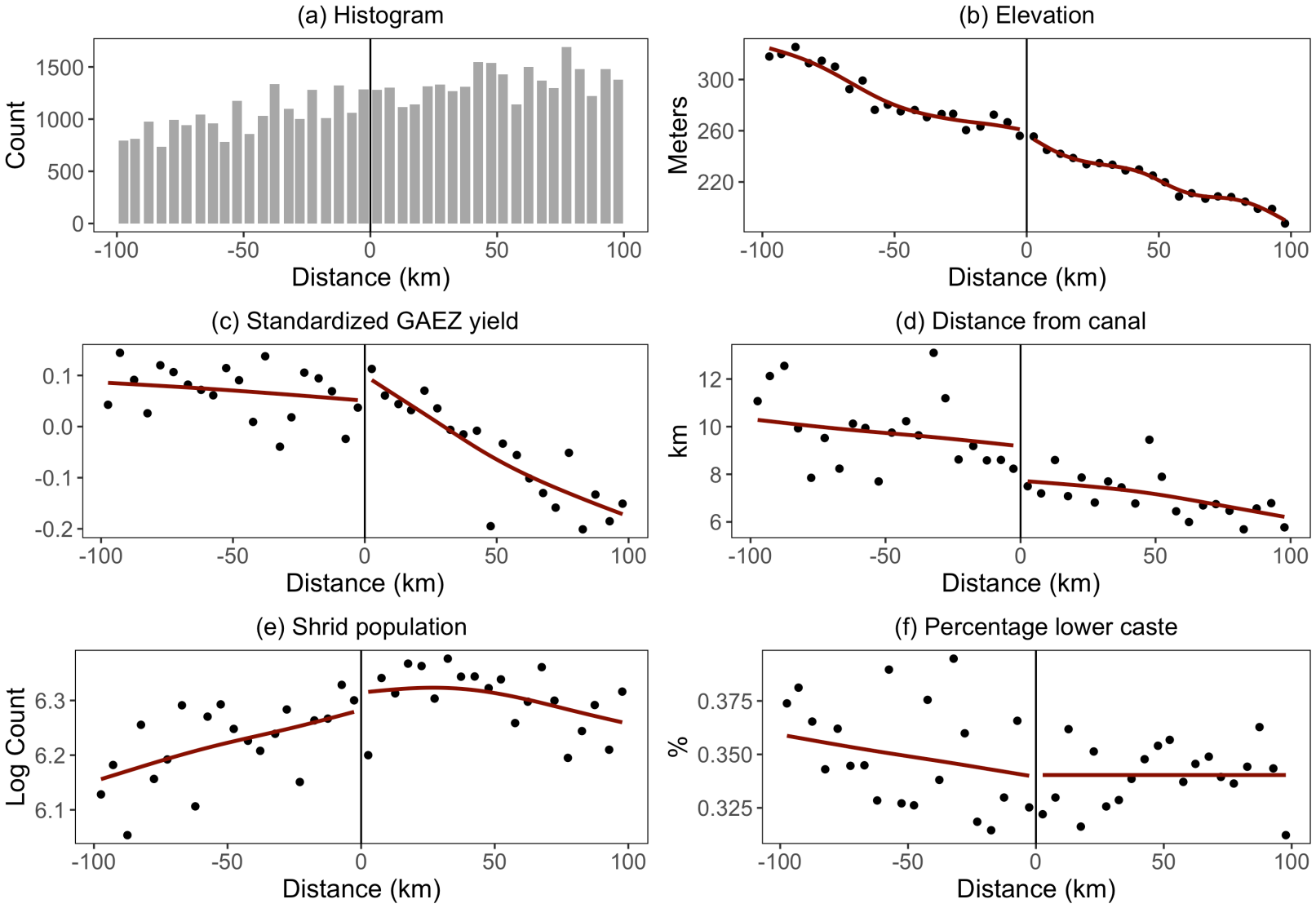

Figure 4: Continuity tests of a selection of covariates. The x-axis is distance along a river relative to a heavily-polluting industrial site. Areas with positive distance are downstream of the site; negative distance is upstream. Dots are binned scatterplots, showing means of each variable within quantiles of the running variable, after partialing out site fixed effects. Lines are cubic splines fitted separately on each side of the graph. Graphs illustrate relationships visually; statistical inference is left for the regressions. ↩

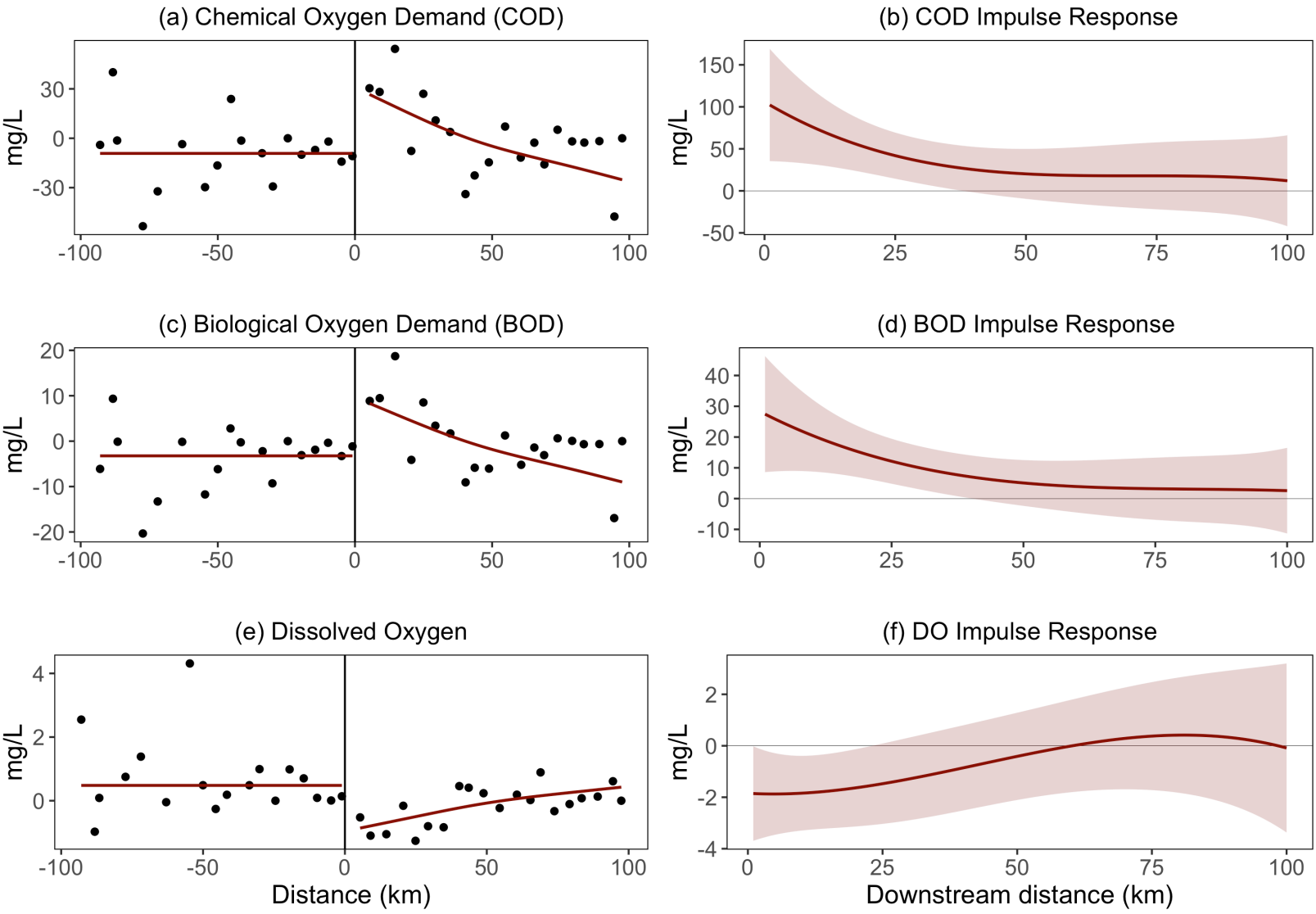

Figure 5: Regression discontinuity plots for pollution measurements. Graphs on the left plot mean values of each parameter within quantile bins of distance from a heavily-polluting industrial site; each dot represents approximately 260 observations. Positive distance values indicate that the monitoring station is downstream of the industrial site; negative values are upstream stations. Values are log-transformed and adjusted for site-by-year fixed effects before binning. Fitted cubic splines illustrate overall patterns. Graphs on the right plot estimated impulse response functions (with 95% confidence intervals), showing how pollution concentrations decay downstream of an industrial site. These functions are cubic splines estimated relative to the upstream linear trend. ↩

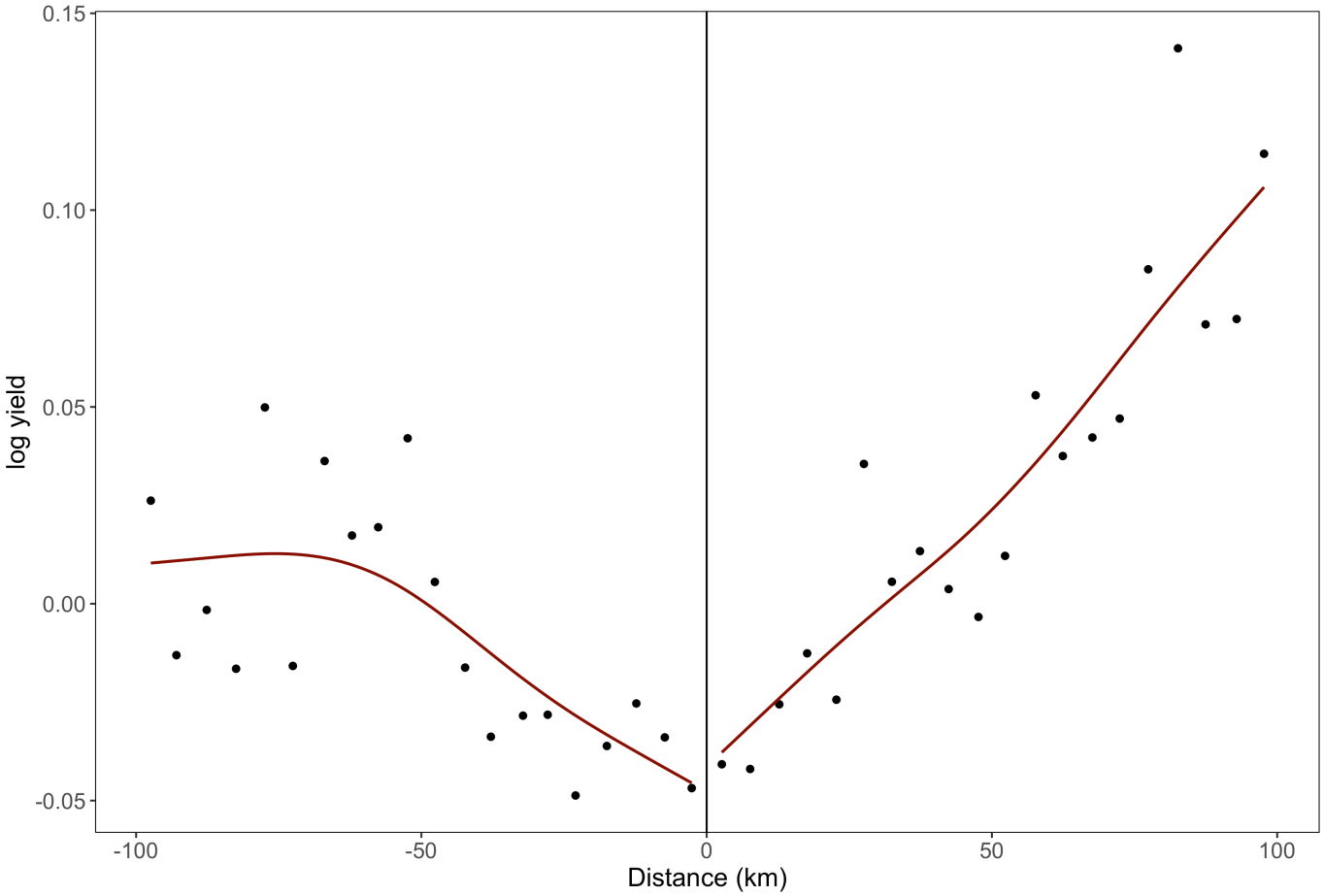

Figure 6: Regression discontinuity plots for predicted agricultural yield. Procedure to calculate predicted yield is described in section 4 and Table 1. The x-axis is distance along a river relative to a heavily-polluting industrial site. Areas with positive distance are downstream of the site; negative distance is upstream. Dots are binned scatterplots, showing means of each variable within quantiles of the running variable, after partialing out site fixed effects. Lines are cubic splines fitted separately on each side of the graph. Graphs illustrate relationships visually; statistical inference is left for the regressions. ↩

11 Tables

Table 1: Correlation of Satellite-based Proxies with District Agricultural Output

Dependent Variable: *log(Revenue Value of Yield)*					
Explanatory Variables	(1)	(2)	(3)	(4)	(5)
Intercept	10.0	9.42	9.36	8.20	8.48
	(0.030)	(0.024)	(0.030)	(0.161)	(0.096)
log(Max VI - Min VI)	0.661	0.058	0.443	0.102	
	(0.044)	(0.005)	(0.045)	(0.012)	
Max NDVI					-4.89
					(0.639)
Min NDVI					0.575
					(1.57)
Max EVI					1.45×10^{-7}
					(1.09×10^{-8})
Min EVI					-8.16×10^{-5}
					(0.0001)
Max NDVI705					10.1
					(0.851)
Min NDVI705					1.60
					(1.35)
Max NDVI740					-4.68
					(1.34)
Min NDVI740					0.704
					(1.33)
Max GCVI					0.001
					(0.001)
Min GCVI					0.027
					(0.437)
Max MTCI					-1.09×10^{-7}
					(6.79×10^{-8})
Min MTCI					-1.22×10^{-7}
					(7×10^{-8})
Vegetation Index (VI)	NDVI	EVI	GCVI	MTCI	
Observations	1,371	1,371	1,371	1,371	1,371
R2	0.205	0.076	0.187	0.064	0.390

Notes: Predictive modesl of observed crop yields (in district-level aggregate data) with respect to satellite-based measures of agricultural production. Coefficients are estimated from regressions of log crop revenue per hectare on remote sensing measures without any fixed effects. Vegetation indices are calculated at pixel-level in Google Earth Engine (GEE) using a cropland mask. Columns 1-4 include the district mean of the log of each pixel's difference between maximum and minimum VI values within a year. Column 5 includes the district mean of the maximum and minimum values for all VIs together. Standard errors (in parentheses) are clustered by district. ↩

Table 2: Summary Statistics

Variable	Mean	SD	N
Panel A: Pollution			
Dissolved Oxygen (mg O_2/l)	6.39	1.97	2299
Chemical Oxygen Demand (mg O_2/l)	43.39	78.40	1979
Biological Oxygen Demand (mg O_2/l)	10.06	20.42	2342
Panel B: Agricultural Output			
Predicted log yield (Rs/ha)	1.20	1.34	100020
Panel C: Agricultural Inputs			
Crop Area under Cultivation per capita (ha)	12.58	82.38	59568
Share of Employment in Ag	0.70	0.22	59800
Share of Crop area under Irrigation	0.51	0.39	59800
Any Irrigation from Rivers (=1)	0.08	0.27	59800
Any Irrigation from Canals (=1)	0.30	0.46	59800
Any Irrigation from Wells (=1)	0.75	0.43	59800

Notes: Summary statistics for the full sample of villages that are either upstream or downstream of severely-polluting industrial sites. Pollution data come from laboratory tests of samples taken at water quality monitoring stations maintained by the Central Pollution Control Board. Predicted yield is calculated using the estimated model in column 5 of Table 1 applied to village-level vegetation indices calculated from Google Earth Engine. Agricultural inputs come from the Population Census of 2001. ↩

Table 3: RD Estimates for Pollution

		RD Bandwidth		
Dependent Variable	Mean [SD]	[25 km]	[50 km]	[100 km]
Biological Oxygen Demand	10.064	23.2	21.3	19.2
	[20.417]	(13.9)	(9.8)	(6.1)
Observations		965	1,549	2,342
R2		0.71	0.62	0.50
Chemical Oxygen Demand	43.388	79.6	80.6	67.9
	[78.399]	(44.0)	(32.0)	(20.2)
Observations		811	1,302	1,979
R2		0.74	0.71	0.62
Dissolved Oxygen	6.388	-1.1	-1.8	-2.0
	[1.965]	(0.80)	(0.57)	(0.44)
Observations		932	1,507	2,299
R2		0.84	0.75	0.67
Distance		X	X	X
Distance X Downstream		X	X	X
Industry X Year FE		X	X	X

Notes: Estimated effects of severely-polluting industrial sites on water pollution concentrations in nearby rivers, immediately downstream of the sites. Dependent variables are listed in rows. Column 2 of the table presents the mean and standard deviation (in brackets) of the dependent variable for the 100 km bandwidth. Each cell in columns 3-5 reports the estimated coefficient on the Downstream indicator variable, controlling linearly for distance on each side of the industrial site along with site-by-year fixed effects. Observations are limited to monitoring stations within the specified bandwidth of the industrial site and are weighted using a triangular kernel. Standard errors clustered by village are provided in parenthesis in columns 3-5. ↩

Table 4: RD Estimates for Predicted Yield

Dependent variable: *Predicted Log Revenue Value of Yield*			
	RD Bandwidth		
	[25 km]	[50 km]	[100 km]
Panel A: Main RD Estimate			
Downstream effect	-0.007	0.009	-0.005
	(0.010)	(0.008)	(0.025)
Observations	25,292	51,929	100,020
Panel B: Robustness to controling for distance to river			
Downstream effect	-0.007	0.009	-0.006
	(0.010)	(0.008)	(0.025)
Observations	25,292	51,929	100,020
Panel C: Robustness to controling for pre-treatment variables			
Downstream effect	-0.017	-0.005	-0.024
	(0.010)	(0.007)	(0.035)
Observations	25,289	51,926	100,014
Panel D: Robustness to controling for irrigation dummies			
Downstream effect	-0.007	0.010	-0.004
	(0.010)	(0.008)	(0.024)
Observations	25,292	51,929	100,020
Distance	X	X	X
Distance X Downstream	X	X	X
Industry X Year FE	X	X	X

Notes: Estimated effects of severely-polluting industrial sites on predicted yield in villages immediately downstream of the sites. Each cell reports the estimated coefficient on the Downstream indicator variable, controlling linearly for distance on each side of the industrial site along with site-by-year fixed effects. Sample includes villages within 20 km of a flow path that passes near each industrial site, as defined in the text. Panel A presents the main RD estimates while the other panels present robustness results. Observations are limited to villages within the specified bandwidth of the industrial site and are weighted by village crop area multiplied by a triangular kernel. Standard errors (in parentheses) are clustered by village. ↩

Table 5: RD Estimates for Agricultural Inputs

		RD Bandwidth		
Dependent Variable	Mean [SD]	[25 km]	[50 km]	[100 km]
Share of Employment in Ag	0.725	0.006	-0.017	-0.012
	[0.211]	(0.011)	(0.008)	(0.010)
Crop Area under Cultivation per capita	48.562	0.214	-0.304	6.11
	[170.754]	(0.431)	(0.247)	(14.8)
Share of Crop area under Irrigation	0.546	0.034	0.028	0.0008
	[0.4]	(0.011)	(0.008)	(0.020)
Any Irrigation from Rivers	0.067	-0.069	0.002	0.030
	[0.25]	(0.019)	(0.015)	(0.022)
Any Irrigation from Canals	0.348	-0.105	-0.034	0.025
	[0.476]	(0.025)	(0.019)	(0.025)
Any Irrigation from Wells	0.756	0.021	0.021	0.041
	[0.429]	(0.018)	(0.013)	(0.011)
Observations		12,121	24,641	47,932
Distance		X	X	X
Distance X Downstream		X	X	X
Industry FE		X	X	X

Notes: Estimated effects of severely-polluting industrial sites on agricultural inputs in villages immediately downstream of the sites. Dependent variables are listed in rows. Column 2 of the table presents the mean and standard deviation (in brackets) of the dependent variable for the 100 km bandwidth. Each cell in columns 3-5 reports the estimated coefficient on the Downstream indicator variable, controlling linearly for distance on each side of the industrial site along with site-by-year fixed effects. Sample includes villages within 20 km of a flow path that passes near each industrial site, as defined in the text. Observations are limited to villages within the specified bandwidth of the industrial site and are weighted by village crop area multiplied by a triangular kernel. Standard errors (in parentheses) are clustered by village. ↩

12 Appendix Tables

Table 6: RD Estimates for Continuity of Covariates

		RD Bandwidth		
Dependent Variable	Mean [SD]	[25 km]	[50 km]	[100 km]
Panel A: Physical Characteristics				
Distance from canal (km)	8.523 [11.439]	-0.333 (1.10)	0.122 (1.07)	-0.725 (0.976)
Distance from nearest town (km)	80.99 [735.755]	-3.10 (1.02)	-3.18 (0.913)	-0.753 (1.10)
Elevation (m)	264.456 [170.14]	-4.11 (3.72)	-7.53 (3.00)	-9.74 (5.49)
Panel B: GAEZ potential yield - High Input Scenario (kg/ha)				
Chickpea	0.594 [0.513]	-0.018 (0.027)	-0.022 (0.022)	0.0004 (0.027)
Cotton	0.771 [0.165]	0.003 (0.014)	0.006 (0.013)	0.005 (0.010)
Dryland rice	1.17 [1.218]	0.051 (0.023)	0.054 (0.027)	0.051 (0.033)
Gram	1.474 [0.412]	0.012 (0.031)	0.017 (0.027)	0.013 (0.026)
Groundnut	1.393 [0.502]	0.021 (0.029)	0.016 (0.026)	0.015 (0.024)
Maize	6.735 [1.939]	0.046 (0.135)	0.085 (0.116)	0.104 (0.116)
Pearl millet	1.361 [1.29]	0.035 (0.028)	0.045 (0.030)	0.061 (0.040)
Pigeon pea	1.917 [0.639]	0.026 (0.040)	0.025 (0.036)	0.024 (0.034)
Rapeseed	0.858 [0.645]	0.013 (0.021)	0.010 (0.018)	0.025 (0.017)
Sorghum	5.931 [1.251]	0.007 (0.113)	0.057 (0.099)	0.104 (0.091)
Soybean	2.127 [0.767]	0.042 (0.049)	0.038 (0.044)	0.026 (0.041)

continued

Table 6: RD Estimates for Continuity of Covariates (Continued)

Sugarcane	1.166 [1.679]	0.045 (0.027)	0.087 (0.035)	0.113 (0.062)
Sunflower	1.035 [0.752]	0.002 (0.029)	-0.067 (0.047)	-0.091 (0.060)
Wetland rice	1.717 [1.061]	0.031 (0.042)	0.013 (0.042)	0.043 (0.056)
Wheat	1.307 [1.131]	0.018 (0.036)	0.011 (0.027)	0.034 (0.029)
Normalized All Crops	-0.281 [0.777]	0.024 (0.049)	0.023 (0.044)	0.031 (0.038)
Panel C: Amenities: Facility Available in Village? (1 = yes, 0 = no)				
Banking	0.152 [0.359]	-0.033 (0.020)	-0.024 (0.019)	-0.011 (0.016)
Communication	0.566 [0.496]	0.013 (0.030)	-0.010 (0.025)	0.006 (0.019)
Medical	0.539 [0.499]	-0.023 (0.031)	-0.026 (0.027)	-0.003 (0.026)
Postal	0.682 [0.466]	0.008 (0.026)	0.016 (0.021)	0.044 (0.018)
Paper and magazines	0.655 [0.476]	-0.073 (0.045)	-0.027 (0.029)	0.014 (0.024)
Educational	0.917 [0.276]	-0.006 (0.010)	-0.004 (0.008)	0.007 (0.010)
Drinking water	0.998 [0.043]	-0.002 (0.003)	-0.002 (0.004)	-0.001 (0.003)
Panel D: Social and Demographic Characteristics				
Household size	5.764 [0.873]	0.073 (0.045)	0.045 (0.037)	0.022 (0.048)
Literacy Rate (percent)	0.504 [0.14]	-0.004 (0.011)	-0.0005 (0.008)	-0.0005 (0.009)
Log Village Area	6.281 [1.056]	-0.112 (0.055)	-0.067 (0.054)	-0.009 (0.050)
Log Population	7.389 [1.082]	-0.109 (0.059)	-0.084 (0.062)	-0.029 (0.057)
Share of Scheduled Caste/Tribe Population	0.307	-0.022	-0.007	-0.008

continued

Table 6: RD Estimates for Continuity of Covariates (Continued)

	[0.245]	(0.023)	(0.018)	(0.015)
Observations		12,103	24,421	47,624
Distance		X	X	X
Distance X Downstream		X	X	X
Industry FE		X	X	X

Notes: Tests of continuity in river space at severely-polluting industrial sites, for covariates that are either fixed in time or unlikely to be affected by the presence of industrial pollution. Dependent variables are listed in rows. Column 2 of the table presents the mean and standard deviation (in brackets) of the dependent variable for the 100 km bandwidth. Each cell in columns 3-5 reports the estimated coefficient on the Downstream indicator variable, controlling linearly for distance on each side of the industrial site along with site-by-year fixed effects. Sample includes villages within 20 km of a flow path that passes near each industrial site, as defined in the text. Observations are limited to villages within the specified bandwidth of the industrial site and are weighted by village crop area multiplied by a triangular kernel. Standard errors (in parentheses) are clustered by village. ↩

Table 7: RD Estimates for other measures of Pollution

		RD Bandwidth		
Dependent Variable	Mean [SD]	[25 km]	[50 km]	[100 km]
Calcium (mg/L)	88.556	96.3	114.1	86.2
	[103.896]	(96.4)	(61.1)	(37.1)
Observations		771	1,208	1,818
Chloride (mg/L)	151.152	88.0	424.3	337.8
	[441.251]	(378.3)	(338.3)	(235.2)
Observations		799	1,260	1,887
Hardness (mg/L)	180.486	191.7	249.0	183.6
	[203.994]	(204.2)	(136.3)	(85.7)
Observations		801	1,260	1,892
Magnesium (mg/L)	52.481	40.8	50.4	36.9
	[58.206]	(38.3)	(26.5)	(17.6)
Observations		760	1,190	1,790
Nitrate (mg/L)	0.982	0.323	0.417	0.414
	[1.121]	(0.297)	(0.192)	(0.140)
Observations		228	383	582
Nitrite (mg/L)	0.489	0.047	0.157	0.195
	[0.958]	(0.073)	(0.141)	(0.120)
Observations		211	361	555
pH	7.687	-0.223	-0.465	-0.401
	[0.534]	(0.257)	(0.248)	(0.178)
Observations		995	1,581	2,384
Potassium (mg/L)	6.782	-8.39	21.0	16.7
	[18.605]	(19.4)	(14.1)	(10.2)
Observations		105	177	268
Sodium (mg/L)	163.909	128.7	597.2	464.0
	[475.09]	(535.9)	(445.6)	(298.5)
Observations		584	903	1,375
Sulphate (mg/L)	79.096	197.4	249.3	176.1
	[204.296]	(153.3)	(121.0)	(82.3)
Observations		784	1,232	1,845
Total Dissolved Solids (mg/L)	684.531	835.4	1,880.6	1,483.8
	[1507.236]	(1,394.2)	(1,116.1)	(776.8)
Observations		689	1,104	1,669
Total Fixed Solids (mg/L)	578.268	394.0	1,703.1	1,423.2

continued

Table 7: RD Estimates for other measures of Pollution (Continued)

	[1333.988]	(1,347.1)	(1,196.7)	(893.4)
Observations		537	884	1,315
Total Suspended Solids (mg/L)	83.763	47.2	102.7	80.8
	[119.596]	(109.2)	(56.5)	(32.4)
Observations		181	302	475
Fecal coliform (CFU/100 ml)	3309110.33	12,051,694.0	10,865,867.7	8,661,662.6
	[74525362.027]	(15,480,806.9)	(9,744,037.8)	(7,822,238.4)
Observations		828	1,345	1,963
Total Coliform (CFU/100 ml)	3974308.496	13,255,235.8	12,737,779.5	10,209,250.2
	[73801861.013]	(17,491,647.2)	(11,311,778.3)	(9,094,242.0)
Observations		847	1,361	2,038
Turbidity (NTU)	52.106	8.07	16.3	23.1
	[67.796]	(12.6)	(10.4)	(8.31)
Observations		736	1,166	1,725
Distance		X	X	X
Distance X Downstream		X	X	X
Industry X Year FE		X	X	X

Notes: Estimated effects of severely-polluting industrial sites on water pollution concentrations in nearby rivers, immediately downstream of the sites. Dependent variables are listed in rows. Column 2 of the table presents the mean and standard deviation (in brackets) of the dependent variable for the 100 km bandwidth. Each cell in columns 3-5 reports the estimated coefficient on the Downstream indicator variable, controlling linearly for distance on each side of the industrial site along with site-by-year fixed effects. Observations are limited to monitoring stations within the specified bandwidth of the industrial site and are weighted using a triangular kernel. NTU is Nephelometric Turbidity Units. CFU is Colony Forming Units. pH is measured in base-10 log units. Standard errors (in parentheses) are clustered by village. ↩

Table 8: RD heterogeneity by distance from river

	Distance from river bin			
RD Bandwidth	[0-5 km]	[5-10 km]	[10-15 km]	[15-20 km]
25 km	-0.026	0.021	-0.047	-0.006
	(0.016)	(0.019)	(0.019)	(0.017)
Observations	5,873	6,040	6,542	6,837
R2	0.80	0.75	0.66	0.79
50 km	0.003	0.033	-0.009	0.00006
	(0.013)	(0.014)	(0.014)	(0.011)
Observations	12,301	12,780	13,206	13,642
R2	0.58	0.75	0.64	0.75
100 km	-0.075	0.026	0.021	0.025
	(0.086)	(0.010)	(0.014)	(0.011)
Observations	24,767	24,841	24,923	25,489
R2	0.03	0.72	0.61	0.72
Distance	X	X	X	X
Distance X Downstream	X	X	X	X
Industry X Year FE	X	X	X	X

Notes: Estimated effects, by distance from the river, of severely-polluting industrial sites on predicted yield in villages immediately downstream of the sites. Dependent variable is always the predicted revenue value of yield. Each cell reports the estimated coefficient from a separate regression of yield on the Downstream indicator variable, controlling linearly for distance on each side of the industrial site along with site-by-year fixed effects. Each row limits the RD sample to villages within the specified bandwidth of the industrial site. The sample for each column contains villages falling within the specified distance bin of a flow path that passes near each industrial site, as defined in the text. Regressions are weighted by village crop area multiplied by a triangular kernel. Standard errors (in parentheses) are clustered by village. ↩

Table 9: RD heterogeneity by highly polluting industry employment share

	HPI classification	
RD Bandwidth	[Below median]	[Above median]
25 km	-0.013	0.001
	(0.013)	(0.013)
Observations	12,521	12,771
R2	0.72	0.72
50 km	0.005	0.013
	(0.010)	(0.009)
Observations	25,523	26,406
R2	0.62	0.70
100 km	0.019	-0.027
	(0.008)	(0.045)
Observations	49,985	50,035
R2	0.60	0.04
Distance	X	X
Distance X Downstream	X	X
Industry X Year FE	X	X

Notes: Estimated effects, by total employment in highly polluting industries, of severely-polluting industrial sites on predicted yield in villages immediately downstream of the sites. Dependent variable is always the predicted revenue value of yield. Each cell reports the estimated coefficient from a separate regression of yield on the Downstream indicator variable, controlling linearly for distance on each side of the industrial site along with site-by-year fixed effects. Sample includes villages within 20 km of a flow path that passes near each industrial site, as defined in the text. Each row limits the RD sample to villages within the specified bandwidth of the industrial site. The sample for each column contains villages falling within above or below median total employment in highly polluting industries across the 48 severely polluted sites. Regressions are weighted by village crop area multiplied by a triangular kernel. Standard errors (in parentheses) are clustered by village. ↩

Table 10: RD heterogeneity by irrigation source

	Irrigation Available from Source?	
RD Bandwidth	[No]	[Yes]
Panel A: Canals		
25 km	-0.007	-0.019
	(0.011)	(0.022)
50 km	0.008	-0.001
	(0.009)	(0.014)
100 km	-0.009	-0.015
	(0.039)	(0.017)
Panel B: Wells		
25 km	-0.039	-0.007
	(0.014)	(0.011)
50 km	-0.007	0.007
	(0.013)	(0.009)
100 km	-0.009	0.018
	(0.040)	(0.011)
Panel C: Rivers		
25 km	-0.007	-0.037
	(0.010)	(0.035)
50 km	0.008	0.014
	(0.008)	(0.027)
100 km	-0.008	0.050
	(0.027)	(0.021)
Distance	X	X
Distance X Downstream	X	X
Industry X Year FE	X	X

Notes: Estimated effects, by presence of irrigation source, of severely-polluting industrial sites on predicted yield in villages immediately downstream of the sites. Dependent variable is always the predicted revenue value of yield. Each cell reports the estimated coefficient from a separate regression of yield on the Downstream indicator variable, controlling linearly for distance on each side of the industrial site along with site-by-year fixed effects. Sample includes villages within 20 km of a flow path that passes near each industrial site, as defined in the text. Each row limits the RD sample to villages within the specified bandwidth of the industrial site. The sample for each column contains villages by the presence of irrigation source. Regressions are weighted by village crop area multiplied by a triangular kernel. Standard errors (in parentheses) are clustered by village. ↩

Table 11: RD Estimates for District-level Actual Yield

	RD Bandwidth		
Dependent Variable	[25 km]	[50 km]	[100 km]
Log Revenue Value	-0.280	-0.173	-0.003
	(0.187)	(0.195)	(0.089)
Observations	2,260	4,018	7,392
R2	0.94	0.88	0.79
Log Revenue	0.227	0.229	0.078
	(0.213)	(0.219)	(0.154)
Observations	1,954	3,484	6,393
R2	0.96	0.90	0.83
Distance	X	X	X
Distance X Downstream	X	X	X
Sample Share	X	X	X
Industry X Year FE	X	X	X

Notes: Regressions report the downstream effect on each outcome variable in aggregate district-level data. Districts may contain areas of land both upstream and downstream of polluting sites, as well as areas that do not fall within our analytical sample at all (neither upstream nor downstream). To approximate an RD design as closely as possible, we estimate regressions of the form $y_{jst} = \beta Downstream_{js} + \phi Sample_{js} + \gamma Distance_{js} + \delta Distance_{js} \times Downstream_{js} + \alpha_{st} + \varepsilon_{jst}$. Here, the treatment variable $Downstream_{js}$ is the proportion of land within each district that falls within the downstream sample. We control for $Sample_{js}$, the proportion of land that falls within either the downstream or upstream samples. Intuitively, we are asking: For districts with similar amounts of land that fall within our sample, how different is the outcome variable when that land falls downstream of the industrial site? We assume that the parts of each district that do not fall within our sample only contribute noise – their outcomes are uncorrelated with the treatment variable. We continue to control for $Distance_{js}$, the average value of the RD running variable across villages within both upstream and downstream samples, as well as the interaction of average distance with the treatment variable. Standard errors are clustered by village. ↩