Industrial Water Pollution and Agricultural Production in India

Nick Hagerty
Anshuman Tiwari*

29 July, 2022

Abstract

Industrial water pollution is high in many developing countries, but researchers and regulators have paid it less attention than air and domestic water pollution. We estimate the costs of industrial water pollution to agriculture in India, focusing on 71 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 in fact coincide with a large, discontinuous rise in pollutant concentrations in the nearest river. Then, we find that remote sensing measures of crop growth are 2.6 percent lower in villages downstream of polluting sites, relative to villages immediately upstream of the same site in the same year. In terms of agricultural production, this estimate roughly translates to a 1 percent negative decline in crop yields. The effect appears to be driven by reduced yields per cropped land area and not factor reallocation. These results suggest that damages to agriculture may not represent a major cost of water pollution, though many other potential social costs remain unquantified.

*Hagerty: Montana State University (email: nicholas.hagerty@montana.edu); Tiwari: London School of Economics (email: a.tiwari@lse.ac.uk). 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, Esther Duflo, Peter Hull, Simon Jäger, Peiley Lau, 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, Hanna, and Oliva 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 2017).

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 63 industrial sites identified by India's Central Pollution Control Board in 2009 as "severely polluted" with respect to water pollution, out of 88 sites selected for intensive study.

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.

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. We use basic hydrological modeling to define precise upstream/downstream relationships between industrial sites and millions of pixels of satellite imagery. As proxies for agricultural output, we use remote sensing products developed by earth scientists to measure vegetation density, plant health, and metabolic activity. The measure we focus on is the normalized differenced vegetation index (NDVI); we also examine two other measures but find them to be noisier. All three have been shown to reliably predict crop yields across a range of settings (Running et al. 2004; Burke and Lobell 2017; David B. Lobell, Di Tommaso, and Burney 2022); Asher and Novosad (2020); David B. Lobell et al. (2020)]. We also show in our context that NDVI predicts agricultural output in

aggregate statistics.

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 prior levels of pollution in nearby rivers by 140%. Second, we find that remote sensing measures of crop growth are lower downstream of high-polluting industrial sites, but only by 2.6 percent. The estimates are precise; confidence intervals exclude differences of more than 5 percent. A rough conversion implies that the sites reduce true crop yields by about 1 percent, suggesting that even the localized effects of industrial water pollution are small. Third, we document that farmers are neither substituting factors of production away from agriculture nor applying additional compensating inputs. The effects of being downstream on crop area, irrigation, labor, and population are small and statistically insignificant.

Our study focuses on crop yields and does not imply that industrial water pollution is not costly. There are 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 to the existing literature in four ways. First, it provides among the first quasi-experimental evidence on any form of costs of industrial water pollution. In recent papers on India, Do, Joshi, and Stolper (2018) study the effects of industrial water pollution on infant mortality, while Brainerd and Menon (2014) study the effects of agricultural water pollution to child health. In the United States, Keiser and Shapiro (2017) study the effect of all water pollution on property values. Most other economics literature on the costs of water pollution deals with domestic water pollution in the context of providing clean drinking water. Second, this paper documents economic costs of pollution to agriculture; to our knowledge only one other paper does so quasi-

experimentally, but in the context of air pollution Aragón and Rud (2016). Third, it adds to the small but rapidly growing literature on the costs of pollution in developing countries (Jayachandran 2009; Chen et al. 2013; Adhvaryu, Kala, and Nyshadham 2022).

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

2 Background

2.0.1 Industrial 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, Faruqui, and Raschid-Sally 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 con-

tent, and in general, plant height, leaf area, and dry matter to be reduced (World Bank and State Environmental Protection Administration 2007).

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

In contrast with industrial 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.

2.0.2 Remote sensing of crop yields

In order to quantify the effect of industrial water pollution on agricultural output, the ideal data would be at a spatial resolution of tens of meters, similar to the Cropland data layer from the US Department of Agriculture. However, the most granular spatial extent over which Indian agricultural data is collected and reported is the administrative unit of a district, an entity that is about 100 sq km on average. This is far too large for our purposes since water pollution may get diluted over distance; also, within-district sources of pollution may affect only part of the district whereas the agricultural output data are for the district as a whole. We solve this challenge by utilizing agricultural proxies from remote sensing data that have been widely used in the scientific agronomic

literature to measure crop yields (Running et al. 2004; Burke and Lobell 2017; David B. Lobell, Di Tommaso, and Burney 2022), and are increasingly common within the economics literature as well (Asher and Novosad 2020; David B. Lobell et al. 2020).

The two most commonly used and related vegetation indices (VIs) to proxy for agricultural yields are the Normalized Difference in Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI). These indices aim to capture the amount of photosynthetic activity in plants in the following way. The chlorophyll 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 same 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 the Chlorophyll pigment will reflect less red light and more near-infrared light. The NDVI provides a simple mathematical formula to compare these two reflectance values and thereby establish the strength of photosynthetic activity in plant matter, independent of other land cover. EVI is very 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). NDVI is strictly bound by the interval [-1, 1] whereas the EVI may have values outside that range, although in practice this is rare.

These measures are both quite effective at crop classification tasks (Wardlow and Egbert 2010). NDVI is known to predict well in developed country settings such as wheat yields in Canada (Hochheim and Barber 1998); but importantly NDVI and EVI have also been shown to predict Maize yields in smallholder settings in Kenya (Burke and Lobell 2017), and to do better than farmers' own self-reported yields and at least as well as crop-cutting experiments and (gold standard) full-scale harvests in Eastern Uganda (David B. Lobell et al. 2020). These settings are closer to Indian agriculture where smallholder farms are dominant in overall crop area cultivated. We also confirm that these pixel-level NDVI data, when appropriately aggregated to the district level, strongly predict agricultural yields from administrative data during our sample period (Asher and Novosad 2020).

In addition to these vegetation indices, we also use another satellite-based proxy known as annual Net Primary Production (NPP), an ecological variable that aims to capture total plant biomass (Running et al. 2004).The NPP is based on the idea that the total amount of solar energy absorbed by a plant minus energy lost through growth and maintenance respiration of the plant can be used to measure the amount of carbon per unit area sequestered within living plant biomass. Therefore, it is a measure of the amount of carbon captured by plants in an ecosystem, after accounting for losses due to respiration. Annual NPP has been to shown to correlate with the NDVI (Tucker, Townshend, and Goff 1985). We utilize the NPP as another proxy for agricultural outcomes, and also verify its correlation with district-level administrative data on agricultural yields.

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, Melesse, and Reddi 2016).

3.0.1 Sample selection and treatment definition

The intuition for our research design is to compare agricultural outcomes in villages downstream of heavily-polluting industrial sites to those in villages upstream of the sites. While the basic idea is simple, defining which villages are "upstream" and "downstream" is conceptually challenging.

The first challenge is that to define "downstream" villages, we need to introduce a channel of exposure through which river pollution reaches the villages. Geometrically, the set of points downstream of a point source is a line – the path water follows to reach the ocean – not an area. 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 produces unique spatial patterns of treatment intensity, depending on topography, geology, soils, infrastructure, and irrigation practices.

The second challenge is that there are many plausible ways to define an "upstream" set of villages. One potential definition of "upstream" is the point source's watershed – the land area that drains into that point. But if the point source does not coincide with a river, its watershed may be small or nonexistent (imagine a plant on top of a hill). Another potential definition is the watershed of a nearby river – but which one? Stream networks are fractals and defining a "river" requires choosing an arbitrary threshold. A low threshold may select a minor creek that results in a very small sample of upstream villages. A high threshold may select a river that is far away from the point source, raising the need to trace the path from source to river, as well as the question of how to handle villages in between.

The third challenge is that if the downstream and upstream samples are selected in asymmetric ways, they may not be good counterfactuals for each other. Our goal is to create a single, unified process to select both downstream and upstream villages, despite the inherent geometric asymmetry

of the situation.

Our solution to these challenges 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. First, we follow the site downstream a short distance (25 km, to the upper yellow dot). Second, we follow this point upstream into the uppermost reaches of its watershed (to the lower yellow dot). Third, we find the downstream flow path of this "headwater" point. This headwater flow path forms the base of our analysis. We define our sample as all villages within 25 km of this flow path. We define treatment status by projecting (snapping) villages and the industrial site onto the headwater flow path, and calculating the flow distance between them along this path. Villages are assigned to downstream if their projection is downstream of the projected industrial site and upstream otherwise.

This approach has several advantages. It ensures the upstream and downstream samples are comparable since they are chosen through a unified process. It ensures a substantial sample of upstream villages, since we follow the industrial site's flow path downstream before finding the watershed. By keeping this distance short, we retain the ability to measure effects within short distances of the industrial site. A simpler approach might simply snap the point sources to the nearest major river on a published map and conduct the same upstream-downstream comparison of villages near that river (e.g., He, Wang, and Zhang (2020)). But the nearest major river is not always on the flow path, which may go in a different direction, depending on topography. Our approach reduces measurement error by modeling the actual flow path.

Our approach is agnostic as to the channel of exposure. It captures the average effect of being downstream of a heavily-polluting industrial site, regardless of whether the pollution arrives through rivers, canals, or groundwater. We also extend the main analysis in several ways to try to disentangle these channels of exposure.

The main disadvantage of our approach is an ambiguity in treatment assignment for a narrow range of villages immediately downstream of the projected industrial site. Industrial pollution likely

enters the river (i.e., the headwater flow path) not at the projected industrial site, but rather at the intersection point of the flow paths from the industrial site and the headwater point. Villages in this range include some that are likely affected by pollution (those near the flow path from the industrial site), and also some that are likely unaffected (those on the opposite side of the headwater flow path from the industrial site). Therefore, we plan to also check whether our results are robust to a "donut hole" specification that excludes this set of villages.

3.0.2 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 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. In future work we will implement the optimal bandwidth algorithm of Calonico, Cattaneo, and Farrell (2020). 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 district (the administrative level below state) to account for correlation across space and 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.

To visualize the results, we plot smoothed binscatter graphs. To create these, we first partial out site-by-year fixed effects by regressing y_{ist} on α_{st}, and add the overall sample mean back to the residuals. We then plot means of these values within quantile bins of distance relative to industrial site. We also fit piecewise cubic splines to the values on each side of the graph. To provide fuller context, we show these graphs for distance bandwidths wider than the bandwidths of our regressions. Because the graphs are constructed differently from the regressions, we omit confidence intervals and leave statistical inference for the regression tables.

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. 2016).

3.0.3 Limitations of temporal variation

While our regressions include repeated cross-sections of data, they do not use temporal variation for identification. In principle, using village or monitoring station fixed effects would allow us to control for unobserved time-invariant factors, improving the credibility of our design. Unfortunately, using temporal variation is impractical for several reasons.

One, 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. Two, 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 is known to suffer from data quality limitations (Bardhan 2013). Three, pollution itself cannot be used as an independent variable without an instrument. Pollution concentrations are strongly affected by changes in runoff (as varying volumes of water dilute the same pollution *load*), which itself strongly affects agricultural production through availability of irrigation water.

Last, the timespan of pollution transport is unknown, and we want to capture the effects of pollution through all possible channels. For example, diffusion through groundwater can take years, decades, or more. Using temporal variation would rule out these channels of transport that take longer to operate. We instead estimate the long-term cumulative effects of location relative to highly polluting industrial plants.

4 Data

4.0.1 Sources

4.0.1.1 Industrial sites The Central Pollution Control Board (CPCB) selected 88 industrial sites for detailed, long-term study in 2009. Names of these sites were taken from the CPCB document (Central Pollution Control Board 2009a). We identified the geolocation of each site using Google Earth 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. Details of the scoring methodology are provided in the companion document Criteria for (Central Pollution Control Board 2009b). The CPCB considers a site "severely polluted" if the score for a single pollution type exceeds 50, or if the overall score exceeds 60 (the overall score is a nonlinear combination of the component scores). Sixty-three sites received such a "severe" rating in water pollution in 2009; these constitute our sample.

4.0.1.2 Pollution measurements We use a rich dataset of water pollution measurements along rivers in India collected by the Central Pollution Control Board, as collected and published by Greenstone and Hanna (2014). This dataset includes monthly observations from 459 monitoring stations along 145 rivers between the years 1986 and 2005. We extend this data set by downloading yearly pollution readings for the same stations from 2006-2012 from the Cental Pollution Control Board's website. Then 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 chemical oxygen demand (COD), a standardized laboratory test that serves as an omnibus measure of organic compounds, which industrial plants typically generate in high quantities. Along with the related but narrower test of biochemical oxygen demand (BOD), COD is the Indian government's top priority in regulating industrial wastewater (Duflo et al. 2013). We also report three other widely reported measures: BOD, dissolved oxygen (DO), and electrical conductivity (EC), a measure of salinity.

None of these measures directly measure inorganic pollutants like heavy metals, which, physiologically, are leading candidates for harming crop growth. Unfortunately, heavy metals are measured in less than three percent of observations. Another limitation is that these measures do not exclusively measure industrial pollution; they can also indicate the presence of domestic or municipal pollution (i.e., untreated sewage). Because many industrial sites are located in cities, they may be coincident with domestic pollution, confounding the interpretation of our results as being driven by industrial pollution. However, another consistently measured water quality indicator is fecal coliforms, a count of the number of certain types of bacteria that originate from human waste. Because fecal coliforms are overwhelmingly produced by domestic pollution, not industrial pollution, partialing out fecal coliforms from COD should leave only the component of COD that is not produced by domestic pollution, i.e., industrial pollution. Therefore in some specifications we control for fecal coliforms.

4.0.1.3 Agricultural outcomes via remote sensing

As discussed in the background section, we utilize remote sensing data to construct measures of agricultural outcomes. All of our remote sensing data retrieval and processing are carried out within the Google Earth Engine. For the NDVI and EVI calculations, we utilize data from the Landsat 7 satellite launched by NASA and operated by the US Geological Survey (USGS). The temporal coverage of Landsat 7 from 1999 onward suits our analysis better than the newer Landsat 8 that was launched in 2013. Landsat 7 covers each point on earth every 16 days, thereby providing an image of the globe at that frequency. The spatial res-

olution of 30m of Landsat 7 is also superior to the 250m resolution of the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard NASA's Terra satellite, data from which are also commonly used to calculate these vegetation indices. We utilize the Surface Reflectance product that is recommended for constructing vegetation indices since it includes atmospheric corrections for aerosols, and apply the quality assurance mask that indicates cloud cover over the pixel (Young et al. 2017). We also apply an agricultural land use mask from the Copernicus Global Land Service (CGLS) to ensure that only pixels where agricultural activity is being carried out are included in the sample.

The Net Primary Production measure is non-trivial to calculate. Therefore, we rely on the pre-calculated MOD17A3HGF.006 product based on the MODIS instrument aboard the NASA Terra satellite. These data are available at 500m resolution at an annual frequency, representing the total amount of carbon sequestered in the plant biomass on each pixel. We also apply the agricultural land use mask from CGLS to this product.

We spatially aggregate these pixel-level data to the village-level to match with Population and Economic Census data to conduct our main analysis. Since the NPP is an annual measure for the pixel, we spatially average it to the village-level and then apply the log transform. In order to conduct the aggregation for NDVI and EVI, we first calculate the pixel-level difference in the maximum and minimum values of the two vegetation indices. The idea here is to measure yearly changes in greenness that could be larger for cropland due to the use of inputs, similar to Asher and Novosad (2020). These differenced measures are then log-transformed for easier interpretation. We also utilize the same two indices provided by Asher and Novosad (2020) who did not apply the cropland mask in their calculation.[1] We end up with 5 different remote sensing proxies for agricultural output at the village-level: the Google Earth Engine (GEE) NDVI, EVI and NPP that we calculate, and the SHRUG NDVI and EVI that we download.

[1] Available for download from the SHRUG platform.

4.0.1.4 Agricultural outcomes in aggregate data

We rely on aggregate measures at the district level in India compiled by the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in their District Level Database (DLD). [2] This data contains information on crop area planted, output and prices for all the main crops as well as some peripheral crops. Price data is available for 16 crops, covering about 79% of all area under cultivation. This data contains 571 districts across 20 states from 1990-2015 for the agricultural year that runs from July 1 to June 30.

Our primary outcome of interest is agricultural revenue. To calculate this, we multiply the quantity of each of 16 crops available in the dataset by the 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.

4.0.1.5 Village shapefiles and covariates

Shapefiles for villages and towns in the Population census of 2001 are not publicly available, although covariate data on more than 200 variables indicating employment, population, ameneties and infrastructure are provided on the census website. We download shapefiles cleaned and provided in the 'Indian Village-Level Geospatial Socio-Economic Data Set, v1' by the Socioeconomic Data and Applications Center run by NASA and hosted by Columbia university.[3] These shapefiles come with the various census covariates included in the data. We find village centroids from the polygons.

Boundaries of villages and towns may change over time. Here we rely on the Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG) provided by the Development Data Lab.[4] 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 three population censuses (1991, 2001, 2011). Village/town level administrative data from various censuses and surveys can be linked to each shrid, after aggregating over the appropriate spatial extent. Almost 96% of villages

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

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

[4] Available for download at https://www.devdatalab.org/shrug_download/

from the 2001 population match a single shrid, therefore not needing any spatial aggregation over village polygons or administrative data. For the rest of the villages, we dissolve the polygons boundaries to obtain the shrid boundaries, and aggregate administrative data over the villages within each shrid.

Some baseline village covariates and outcomes are summarized in table 1. We use more covariates to test for continuity at the RD threshold later on. These include, but are not limited to, total population and area of the village; irrigated area (gross irrigated area per gross cropped area, percent); river, canal and groundwater specific measures of the previous variable; and land cropped (net cropped area per total district area, percent).

4.0.2 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 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).

Headwater points. To generate a "headwater" point for each industrial site, we use the Trace Downstream tool in ArcGIS Pro (from the Ready to Use Hydrology toolset) to find the flow path of each industrial site. This flow path is 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. Next, we use the Watershed tool (in the same ArcGIS Pro toolset) 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. We choose the 95th percentile instead of the maximum to avoid erroneous values that sometimes arise at the edges of the watershed polygons.

Sample selection. To define the sample of villages for each industrial site, we find the flow path of each headwater point (again using the Trace Downstream tool), generate a 25-km buffer around each flow path, and intersect this buffer with village centroids. The distance of 25 km represents an estimate of how far away from the river pollution is likely to travel. We use alternative buffer widths in robustness checks.

Village distance and treatment status. To calculate distances for the RD design, we project all village centroids, industrial sites, and monitoring stations into one-dimensional river space. Specifically, we snap all these points to the nearest point along the headwater flow path. 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 the 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.

4.0.3 Continuity tests and summary statistics

We provide summary statistics in Table 1 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 any 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, few coefficients are statistically significant. 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.

5 Results

5.0.1 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. The amount of water pollution released by these sites has not previously been quantified in publicly available sources.

Figure 5 visualizes our main results for pollution. It shows regression discontinuity plots for the four water quality parameters that are both widely reported and associated with industrial pollution: chemical oxygen demand (COD), biological oxygen demand (BOD), dissolved oxygen (DO), and electrical conductivity (EC). These 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 nonparametric curves to show overall patterns.

All four parameters show a discontinuous increase in pollution at the exact location of the industrial sites. COD, BOD, and EC all 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. The shapes of these graphs also show that water pollution dissipates as the river flows downstream. For all four parameters, pollution is highest immediately after the industrial site. It then steadily falls and rejoins the trend implied by the upstream curve at a distance of 100 to 200 km from the industrial site. (The noisy declines after 200 km are likely caused by unmodeled factors or compositional changes in the density of monitoring stations across industrial sites.)

Table 2 quantifies these results. It reports RD estimates from separate regressions 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 0.883 for COD (with a 50-km bandwidth) implies that the average "severely polluted" industrial site increases pollution in nearby rivers by 140% ($e^{0.883} - 1$). Estimated discontinuities are statistically significant for larger bandwidths (50 and 100 km for COD and 100 km for other parameters at a 95% confidence level, as well as 50 km for all parameters at a 90% confidence level). They are not significant for a bandwidth of 25 km, but this is due to a lack of precision. The point estimates remain very similar across bandwidths, while standard errors shrink as bandwidths increase and more data enters the sample.

5.0.2 Agricultural outcomes

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

Figure 6 visualizes our main results for agricultural production. It shows similar RD plots as for pollution, but using observations at the level of village-by-year, instead of station-by-date. None of these plots show an obvious discontinuity at the industrial site. Despite increasing water pollution drastically, industrial sites do not seem to affect downstream vegetation indices, suggesting they do not reduce crop yields. This is true whether we use NDVI or EVI as the outcome measure, or whether we use the cropland-masked indices (from GEE) or the whole-village indices (from SHRUG).

Table 3 quantifies these results. As before, it reports RD estimates for the outcome variables listed in each row for multiple bandwidths (in columns). We show five outcome variables, all log-transformed for a percentage interpretation: cropland-masked NDVI, whole-village NDVI, cropland-masked EVI, whole-village EVI, and net primary productivity (NPP). The NDVI and EVI measures are differenced to adjust for off-season normals; NPP is a cumulative measure based

on the whole year.

Cropland-masked NDVI yields the most precise estimates. (All coefficients represent the effect of a binary treatment variable on a log-unit outcome variable, so their standard errors can be compared directly.) The estimate for a 50-km bandwidth is -0.026, implying that NDVI is 2.6 percent lower immediately downstream of a severely-polluting industrial site. Even clustering conservatively by district, the 95% confidence interval allows us to reject a decrease in NDVI greater than 5 percent, as well as any increase in NDVI. Other measures are less precise but broadly consistent. None of the point estimates is positive, none is larger in magnitude than -0.047, and all 95% confidence intervals exclude reductions of more than 12 percent.

How do these proxies translate to crop yields? We can conduct a back-of-the-envelope calculation using results from Appendix Table 5, in which we regressed district-level agricultural output on the vegetation indices. We found that a 10-percent increase in cropland-masked NDVI predicts a 3.7 percent increase in crop revenues per acre (using state and year fixed effects in row 1, column 2). If this relationship holds equally for all sources of variation in NDVI and revenues, then a decrease in NDVI of 2.6 percent would imply a decrease in crop yields of 1.0 percent.

5.0.3 Agricultural inputs and economic outcomes

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

Table 4, Panel B reports RD estimates for a set of agricultural inputs. Labor, as measured by the share of employment in agriculture, does not change immediately downstream of severely-polluting industrial sites (the point estimate is small and not statistically significant). Neither does land, as measured by crop area under cultivation. Irrigation plausibly might either increase to compensate

for damage from pollution or decrease because the water itself is harmful, but that does not appear to happen overall (share of crop area under irrigation) nor for any particular source (share of irrigation from rivers, canals, or wells).

Table 4, Panel A reports RD estimates for two follow-on economic outcomes. Per capita expenditure does not appear to decrease downstream of industrial sites. There is weak evidence that the rural poverty rate falls, but the magnitudes are small and insignificant for most bandwidths.

6 Explanations

It may be puzzling that near some of the largest point sources of industrial water pollution in the world, crops seem not to be harmed more than 1 to 3 percent. We propose six potential explanations for our results and attempt to evaluate them using available evidence.

Pollution effects are highly localized. One possibility is that the effects of pollution are concentrated in an area too small for even our highly targeted research design to detect. In particular, our analysis includes all villages within 25 km of the flow line; perhaps this radius is too large. In future work, we will test robustness to varying this radius and estimate heterogeneous treatment effects for different width bins around the flow line. However, one implication of this explanation is that if pollution effects are highly localized, the aggregate effects of pollution are probably not very large.

Industrial 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. It is also possible that our estimates, which average over across industrial sites, mask heterogeneity across sites. In future work, we will estimate heterogeneous

treatment effects by site and investigate whether the limited pollution data can shed any light on the differences in pollutants across sites.

Estimates are confounded by beneficial municipal water pollution. Municipal wastewater that is less than completely treated also contains high concentrations of compounds that can serve as fertilizer for crops. (It also can contain disease-causing microorganisms, but these only affect human health, not plant growth). For industrial sites located in cities, our estimates might pick up the effects of municipal wastewater in addition to the effects of industrial pollution. The net effect of both, again, might be near zero. In future work, we will identify cities using population density and fecal coliform measurements and estimate heterogeneous treatment effects by whether the industrial site is colocated with a city.

Pollution harms output quality, not 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. Obvious quality effects may capitalize into prices; other quality effects may not. In future work, we will test for impacts to crop prices in aggregate data, as well as local human health.

Remote sensing proxies are unable to detect yield effects from water pollution. Despite a range of papers in both the economics and scientific literatures that have found satellite-derived vegetation indicies to be useful proxies for crop yields and agricultural output, many questions and uncertainties remain about their capabilities. One possibility is that NDVI and EVI are simply not well-suited to pick up the effects of industrial water pollution on crops. This could be because water pollution affects crops in ways that do not show up in remote sensing measures, although many of the agronomy studies on water pollution do specifically report negative impacts to leaf size and color, characteristics that vegetation indices are well-tailored to measure. It may also be possible that farmers adjust crop choice in response to pollution exposure, even though we do not

see other inputs change. NDVI and EVI 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.[5]

Harm to crop yields truly is small. If none of the previous five explanations is true, then the chief remaining possibility is the simplest one implied by our results: that industrial water pollution does not reduce crop yields very much. Perhaps even the levels of industrial pollution seen in India are not large enough to substantially affect crops. Perhaps the mechanism of exposure is too indirect – since most irrigation water in India is pumped from wells, perhaps industrial effluent filters through enough layers of soil that pollutants are removed or diluted before being taken up by crops.

7 Conclusion

This paper provides the first quasi-experimental evidence on the costs of industrial water pollution to agriculture. We examine 71 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 construct remote sensing proxies for agricultural yields including Normalized Difference Vegetation Index, Enhanced Vegetation Index and Net Primary Production. These proxies have been shows 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 extend the data set on water pollution collected by Greenstone and Hanna (2014) between 1986-2005 to 2012 using yearly data available from the Central Pollution Control Board of India.

[5] Ideally, our analysis would use a dataset like the U.S. Department of Agriculture's Cropland Data Layer to control for crop type, but we are unaware of analogous data for India that is publicly available.

With these data in hand, we conduct our RD analyses. First, we find that the location of these industrial sites coincides with a large, discontinuous jump in water pollution in nearby rivers. Second, we find that each district immediately downstream of these sites has, on average, 1% percent lower crop revenue per hectare (with a 95% confidence interval of -0.2% to -2%) than a corresponding district immediately upstream of the same site, in the same year. Third, we find that this effect is driven by the direct impact on yields; there is no evidence that factor reallocation either mitigates or exacerbates it.

We propose six explanations for these findings. Pollution effects may be highly localized, industrial pollution may have beneficial components for agriculture, municipal pollution that can increase yields may confound the estimates, pollution may harm output quality and not quantity, remote sensing may not be adequate to detect yield effects or finally harms from industrial pollution may truly be small. In future work, we plan to investigate each of these explanations using heterogeneity analyses and further data collection.

8 References

Adhvaryu, Achyuta, Namrata Kala, and Anant Nyshadham. 2022. "Management and Shocks to Worker Productivity." Journal of Political Economy 130 (1): 1–47. https://doi.org/10.1086/717046.

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. https://doi.org/10.1111/ecoj.12244.

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. https://doi.org/10.1111/ecoj.12273.

Asher, Sam, and Paul Novosad. 2020. "Rural Roads and Local Economic Development." American Economic Review 110 (3): 797–823. https://doi.org/10.1257/aer.20180268.

Bardhan, Pranab. 2013. "The State of Indian Economic Statistics: Data Quantity and Quality Issues." University of California, Berkeley. https://eml.berkeley.edu/~webfac/bardhan/papers/EconomicStatistics.pdf.

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. https://doi.org/10.1016/j.jdeveco.2013.11.004.

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. https://doi.org/10.1073/pnas.1616919114.

Bustos, Paula, Bruno Caprettini, and Jacopo Ponticelli. 2016. "Agricultural Productivity and Structural Transformation: Evidence from Brazil." American Economic Review 106 (6): 1320–65. https://doi.org/10.1257/aer.20131061.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2020. "Optimal Bandwidth Choice

for Robust Bias Corrected Inference in Regression Discontinuity Designs." 1809.00236. arXiv.org.

Central Pollution Control Board. 2009a. "Comprehensive Environmental Assessment of Industrial Clusters." December.

———. 2009b. "Criteria for Comprehensive Environmental Assessment of Industrial Clusters." December.

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. https://doi.org/10.1073/pnas.1300018110.

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. https://doi.org/10.1016/j.jdeveco.2018.03.001.

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. https://doi.org/10.1093/qje/qjt024.1.

Duranton, Gilles, Ejaz Ghani, Arti Grover Goswami, and William Kerr. 2016. "A Detailed Anatomy of Factor Misallocation in India." Working {Paper}. Washington, DC: World Bank. https://doi.org/10.1596/1813-9450-7547.

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.

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. https://doi.org/10.3386/w20405.

Ghatak, Maitreesh, and Dilip Mookherjee. 2014. "Land Acquisition for Industrialization and Compensation of Displaced Farmers." Journal of Development Economics 110: 303–12. https://doi.org/10.1016/j.jdeveco.2013.01.001.

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. https://doi.org/10.3390/s16081298.

Greenstone, Michael, and Rema Hanna. 2014. "Environmental Regulations, Air and Water Pollution, and Infant Mortality in India." American Economic Review 104 (10): 3038–72. https://doi.org/10.1257/aer.104.10.3038.

He, Guojun, Shaoda Wang, and Bing Zhang. 2020. "Watering Down Environmental Regulation in China*." The Quarterly Journal of Economics 135 (4): 2135–85. https://doi.org/10.1093/qje/qjaa024.

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. https://doi.org/10.1080/07038992.1998.10874687.

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. https://doi.org/10.1162/qjec.2009.124.4.1403.

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).

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

Demand for Water Quality."

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. https://doi.org/10.1016/j.eiar.2006.08.004.

Lobell, David B., Stefania Di Tommaso, and Jennifer A. Burney. 2022. "Globally Ubiquitous Negative Effects of Nitrogen Dioxide on Crop Growth." Science Advances 8 (22): eabm9909. https://doi.org/10.1126/sciadv.abm9909.

Lobell, David B, 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. https://doi.org/10.1093/ajae/aaz051.

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. https://timesofindia.indiatimes.com/india/indias-88-industrial-clusters-present-a-bleak-picture-of-air-water-and-land-contamination-says-cse-report/articleshow/81221535.cms.

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.

Reddy, V. Ratna, and Bhagirath Behera. 2006. "Impact of Water Pollution on Rural Communities: An Economic Analysis." Ecological Economics 58 (3): 520–37. https://doi.org/10.1016/j.ecolecon.2005.07.025.

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. https://doi.org/10.1641/0006-3568(2004)054%5B0547:ACSMOG%5D2.0.CO;2.

Scott, C. I., N. I. Faruqui, and L. Raschid-Sally. 2004. Wastewater Use in Irrigated Agriculture: Confronting the Livelihood and Environmental Realities. https://doi.org/10.1079/9780851998237.0000.

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. https://doi.org/10.1016/j.agrformet.2014.06.007.

Tucker, Compton J., John R. G. Townshend, and Thomas E. Goff. 1985. "African Land-Cover Classification Using Satellite Data." Science 227 (4685): 369–75. https://doi.org/10.1126/science.227.4685.369.

Vyas, Ananya. 2022. "Explainer: What Is Causing the Mass Death of Fish in India's Water Bodies?" Text. Scroll.in.

Wardlow, Brian D., and Stephen L. Egbert. 2010. "A Comparison of MODIS 250-m EVI and NDVI Data for Crop Mapping: A Case Study for Southwest Kansas." International Journal of Remote Sensing 31 (3): 805–30. https://doi.org/10.1080/01431160902897858.

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

Young, Nicholas E., Ryan S. Anderson, Stephen M. Chignell, Anthony G. Vorster, Rick Lawrence,

and Paul H. Evangelista. 2017. “A Survival Guide to Landsat Preprocessing.” Ecology 98 (4): 920–32. https://doi.org/10.1002/ecy.1730.

9 Figures and Tables

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.) ↩

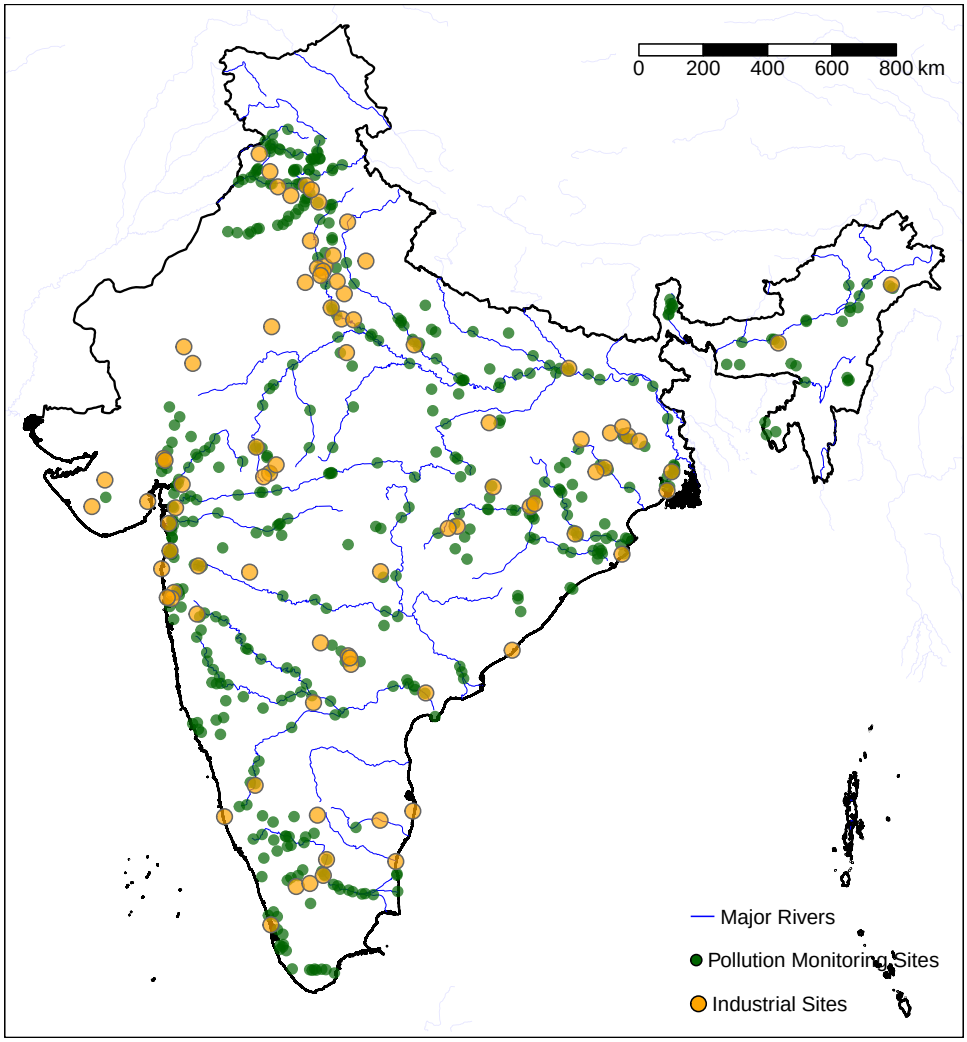

Figure 2: Locations of "severely polluted" industrial sites (orange dots) and water pollution measurement stations (green dots).↩

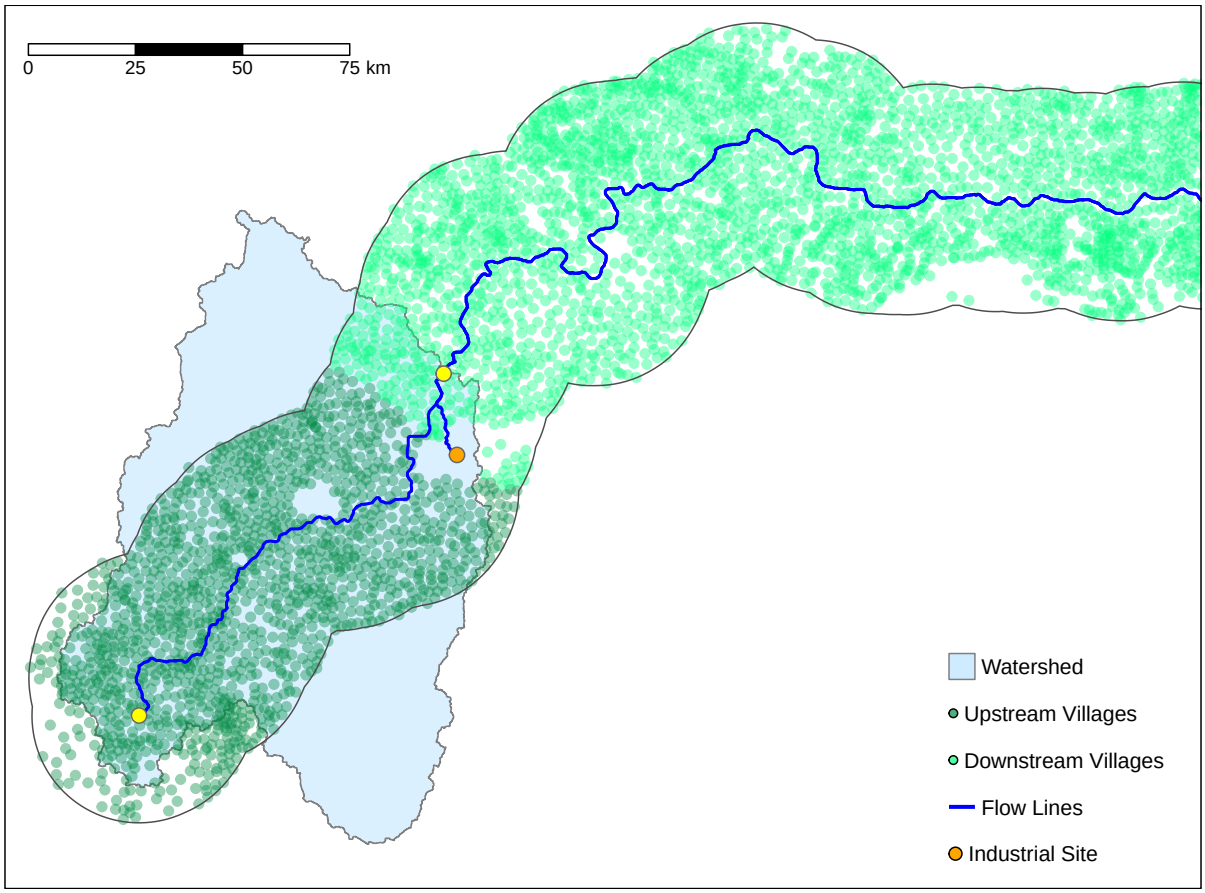

Figure 3: Illustration of the sample selection and treatment assignment for our main research design.
↩

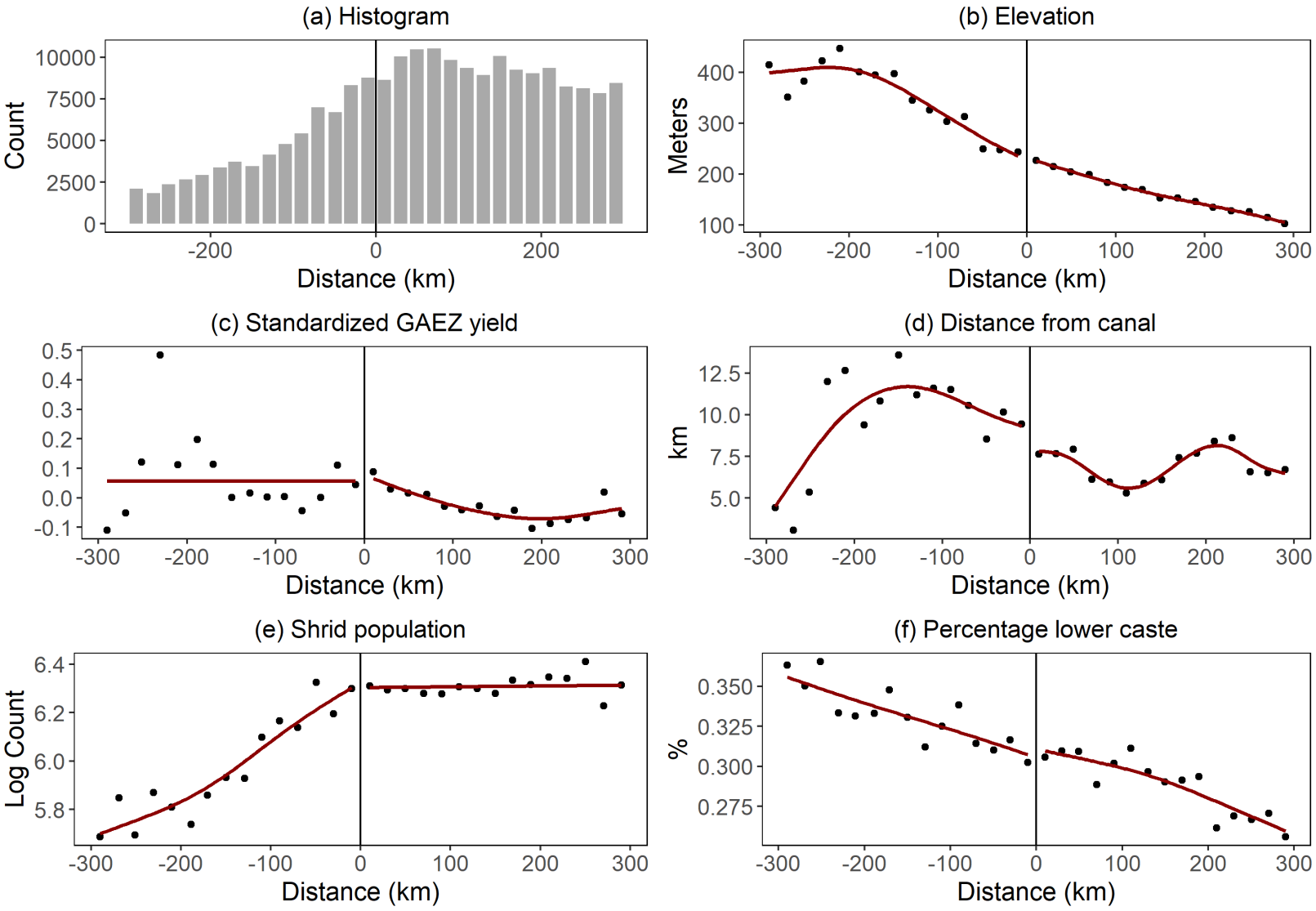

Figure 4: Continuity tests for a selection of covariates. ↩

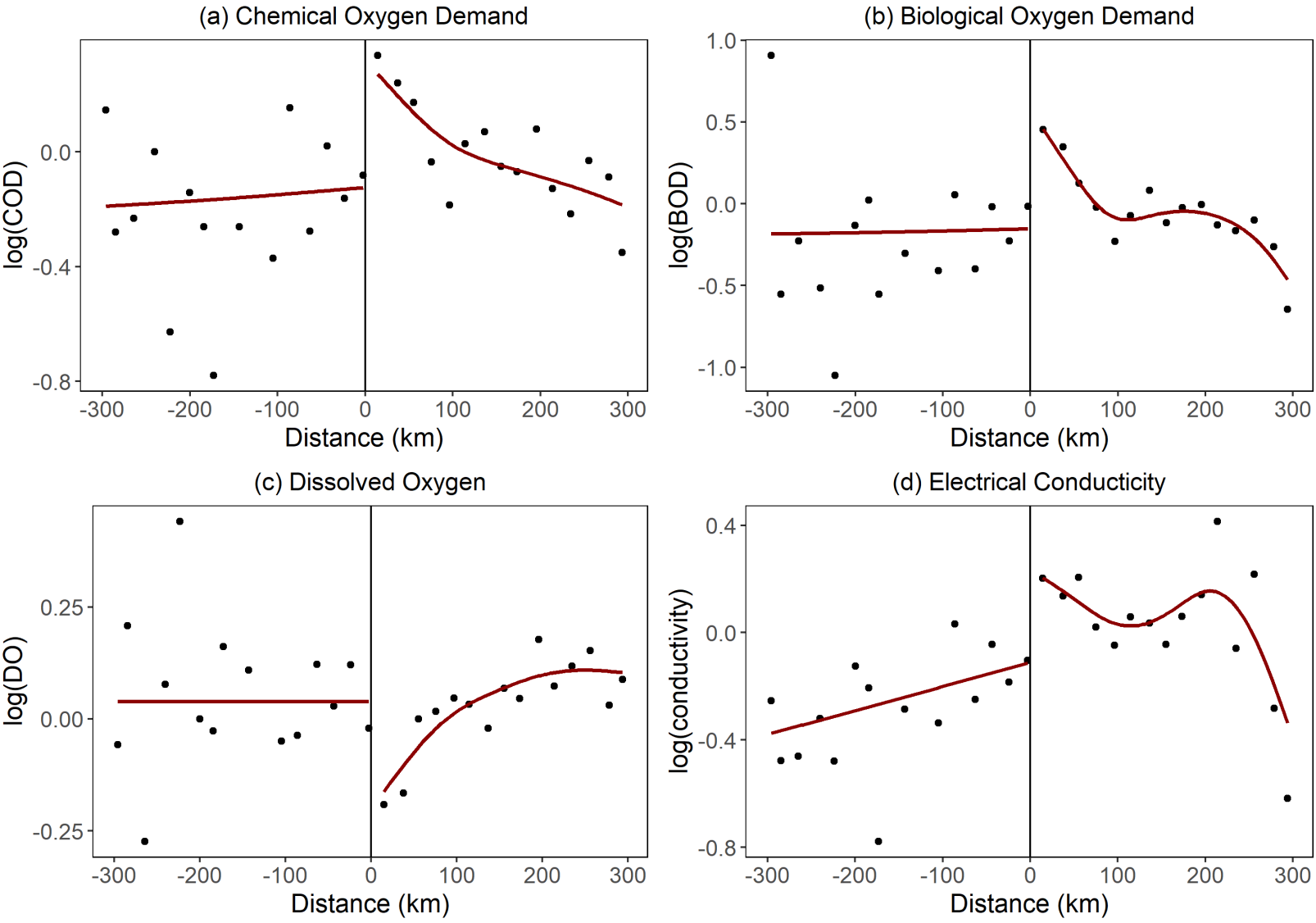

Figure 5: Regression discontinuity plots for pollution measurements. ↩

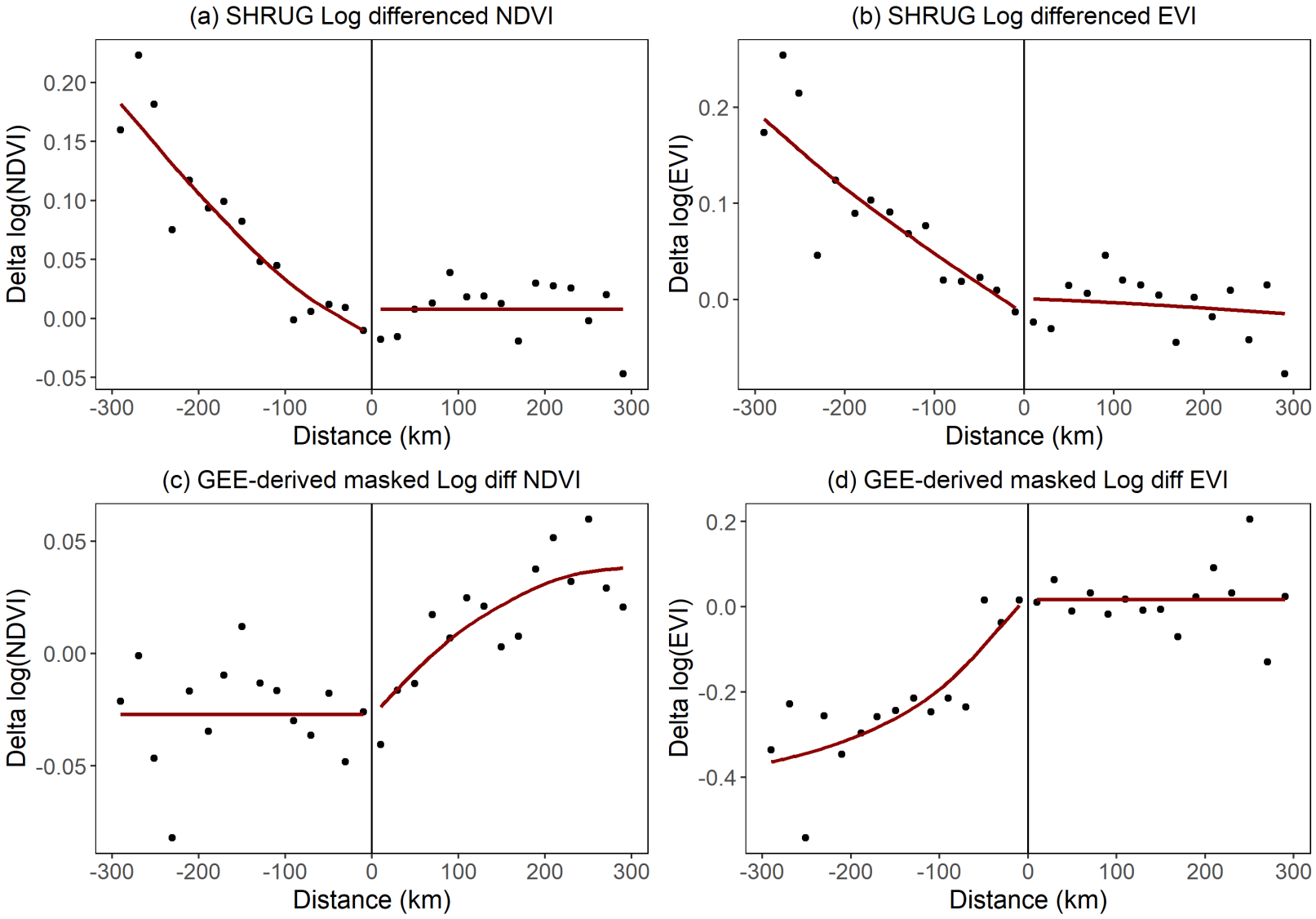

Figure 6: Regression discontinuity plots for measures of agricultural production. ↩

10 Tables

Table 1: Summary Statistics

Variable	Mean	SD	N
Panel A: Pollution			
Dissolved Oxygen (mg O_2/l)	6.46	1.94	3322
Chemical Oxygen Demand (mg O_2/l)	39.73	56.74	2872
Biological Oxygen Demand (mg O_2/l)	8.81	15.26	3414
Electrical Conductivity (millisiemens/cm)	472.02	868.53	3204
Panel B: Agricultural Output			
NDVI (GEE)	-0.81	0.35	1366263
EVI (GEE)	1.64	1.25	1366263
NDVI (SHRUG)	8.12	0.43	1366112
EVI (SHRUG)	7.82	0.51	1366112
Net Primary Production (kg C/m2)	7.47	1.11	1355604
Panel C: Economic Outcomes			
Crop Area under Cultivation per capita (ha)	18.75	106.60	106340
Share of Employment in Ag	0.71	0.21	106829
Share of Crop area under Irrigation	0.52	0.40	106829
Share of Irrigation from Rivers	0.02	0.10	90809
Share of Irrigation from Canals	0.14	0.27	83066
Share of Irrigation from Wells	0.41	0.37	56222
Per Capita Expenditure (Rs)	65.14	1109.76	106829
Rural Poverty Rate	0.26	0.18	103981

continued

Table 1: Summary Statistics

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. NDVI and EVI variables from Google Earth Engine (GEE) are the mean of the log of each pixel's difference between maximum and minimum NDVI values within a year, for all village pixels marked as cropland in the cropland mask. NDVI and EVI variables from SHRUG are the log of the difference between the maximum and early-season village-mean NDVI values within a year, with no cropland mask applied. The net primary production (NPP) variable is the log of the mean of estimated NPP values across all cropland pixels within the village. Economic outcomes come from the Population Census of 2001. ↩

Table 2: RD Estimates for Pollution

	RD Bandwidth		
Dependent Variable	[25 km]	[50 km]	[100 km]
log(Biological Oxygen Demand)	0.977	1.02*	0.909**
	(0.725)	(0.512)	(0.353)
Observations	1,365	2,215	3,414
R2	0.896	0.822	0.742
log(Chemical Oxygen Demand)	0.782	0.883**	0.741**
	(0.520)	(0.416)	(0.300)
Observations	1,137	1,852	2,872
R2	0.842	0.754	0.682
log(Electrical Conductivity)	0.628	0.656*	0.557**
	(0.433)	(0.357)	(0.238)
Observations	1,301	2,105	3,204
R2	0.934	0.922	0.918
log(Dissolved Oxygen)	-0.254	-0.373*	-0.386**
	(0.238)	(0.188)	(0.145)
Observations	1,318	2,145	3,313
R2	0.823	0.714	0.624
Distance	X	X	X
Distance X Downstream	X	X	X
Industry X Year FE	X	X	X

continued

Table 2: RD Estimates for Pollution

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; 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. 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 district. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 3: RD Estimates for Agricultural Outcomes

	RD Bandwidth		
Dependent Variable	[25 km]	[50 km]	[100 km]
log(Differenced GEE NDVI)	-0.017	-0.026**	-0.022*
	(0.013)	(0.013)	(0.013)
Observations	363,062	726,304	1,366,263
R2	0.732	0.706	0.676
log(Differenced SHRUG NDVI)	-0.029	-0.047	-0.046
	(0.028)	(0.030)	(0.030)
Observations	359,128	729,014	1,366,112
R2	0.579	0.586	0.558
log(Differenced GEE EVI)	-0.012	-0.008	-0.008
	(0.045)	(0.042)	(0.050)
Observations	363,062	726,304	1,366,263
R2	0.704	0.686	0.654
log(Differenced SHRUG EVI)	-0.023	-0.042	-0.040
	(0.029)	(0.029)	(0.028)
Observations	359,128	729,014	1,366,112
R2	0.604	0.602	0.572
log(Net Primary Productivity)	-0.0008	-0.010	-0.027
	(0.031)	(0.039)	(0.049)
Observations	360,731	721,243	1,355,604
R2	0.776	0.762	0.730

continued

Table 3: RD Estimates for Agricultural Outcomes

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 remote sensing measures of agricultural production in villages immediately downstream of the sites. Dependent variables are listed in rows; 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 25 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 using a triangular kernel. Standard errors clustered by district. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 4: RD Estimates for Agricultural Inputs and Economic Outcomes

	RD Bandwidth		
Dependent Variable	[25 km]	[50 km]	[100 km]
Panel A: Economic Outcomes			
Per Capita Expenditure	19.2	56.8	62.3
	(36.4)	(75.1)	(71.5)
Rural Poverty Rate	-0.003	-0.013	-0.015*
	(0.011)	(0.009)	(0.009)
Panel B: Agricultural Inputs			
Share of Employment in Ag	0.015	-0.004	0.003
	(0.016)	(0.014)	(0.012)
Crop Area under Cultivation per capita	13.3	36.9	36.1
	(17.9)	(27.1)	(28.2)
Share of Crop area under Irrigation	-0.009	-0.035	-0.028
	(0.034)	(0.044)	(0.044)
Share of Irrigation from Rivers	-0.004	0.003	0.002
	(0.005)	(0.004)	(0.003)
Share of Irrigation from Canals	-0.017	-0.006	-0.007
	(0.018)	(0.013)	(0.015)
Share of Irrigation from Wells	0.041	0.021	0.034
	(0.027)	(0.024)	(0.026)
Observations	11,454	23,740	46,262

continued

Table 4: RD Estimates for Agricultural Inputs and Economic Outcomes

Distance	X	X	X
Distance X Downstream	X	X	X
Industry FE	X	X	X

Notes: Estimated effects of severely-polluting industrial sites on measures of agricultural inputs and economic outcomes in villages immediately downstream of the sites. Regressions are as described in Table 3. *p<0.1; **p<0.05; ***p<0.01. ↩

11 Appendix Tables

Table 5: Correlation of Satellite-based Proxies with Agricultural Output

	GEE				SHRUG			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
log(Differenced NDVI)	0.212***	0.367***	0.020	-0.044***	0.201***	0.192***	0.177***	0.129***
	(0.040)	(0.068)	(0.013)	(0.012)	(0.041)	(0.050)	(0.021)	(0.022)
R2	0.411	0.544	0.746	0.803	0.412	0.529	0.752	0.805
log(Differenced EVI)	0.094***	0.071***	0.059***	-0.012**	0.216***	0.207***	0.183***	0.132***
	(0.013)	(0.018)	(0.006)	(0.005)	(0.036)	(0.043)	(0.021)	(0.022)
R2	0.415	0.523	0.752	0.802	0.421	0.536	0.755	0.806
log(Net Primary Production)	0.007	0.022	-0.019***	0.009**				
	(0.023)	(0.028)	(0.005)	(0.005)				
R2	0.392	0.514	0.746	0.802				
Fixed Effects	State	StateXYear	District	District, Year	State	StateXYear	District	District, Year
Observations	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000

Notes: Predictive elasticities of crop yields (in district-level aggregate data) with respect to satellite-based measures of agricultural production. Coefficients are estimated from district-by-year regressions of log crop revenue per hectare on the remote sensing measures. Remote sensing measures from Google Earth Engine (GEE) apply a cropland mask (columns 1-4); those from the SHRUG database do not (columns 5-8). *p<0.1; **p<0.05; ***p<0.01. ↩

Table 6: RD Estimates for Continuity of Covariates

		RD Bandwidth		
Dependent Variable	Mean [SD]	[25 km]	[50 km]	[100 km]
Panel A: Infrastructure - Facility Available in Village?				
Banking	0.154	-0.033	-0.032	-0.021
	[0.361]	(0.021)	(0.020)	(0.016)
Communication	0.57	-0.004	0.008	0.014
	[0.495]	(0.024)	(0.020)	(0.018)
Medical	0.548	-0.008	-0.025	-0.023
	[0.498]	(0.033)	(0.036)	(0.033)
Postal	0.691	0.004	0.018	0.037**
	[0.462]	(0.021)	(0.015)	(0.015)
Paper and magazines	0.659	-0.083***	-0.015	0.015
	[0.474]	(0.032)	(0.022)	(0.022)
Educational	0.932	-0.003	-0.0004	0.005
	[0.252]	(0.007)	(0.006)	(0.008)
Drinking water	0.998	-0.0002	$1.64times10^{-5}$	0.0002
	[0.048]	(0.002)	(0.002)	(0.002)
Panel B: Physical Characteristics				
Distance from canal (km)	7.849	-0.426	-0.523	-1.24
	[11.585]	(0.833)	(0.761)	(0.794)
Distance from nearest town (km)	52.937	-2.22***	-1.65	-2.15*

continued

Table 6: RD Estimates for Continuity of Covariates

		RD Bandwidth		
Dependent Variable	Mean [SD]	[25 km]	[50 km]	[100 km]
	[549.917]	(0.847)	(1.53)	(1.09)
Elevation (m)	249.425	-3.54	-6.66**	-8.94**
	[169.407]	(3.14)	(3.01)	(4.50)
Panel C: GAEZ potential yield - High Input Scenario				
Normalized All Crops	-0.224	-0.004	-0.008	-0.022
	[0.703]	(0.038)	(0.033)	(0.031)
Chickpea	0.585	-0.025	-0.025	-0.010
	[0.51]	(0.023)	(0.021)	(0.026)
Cotton	0.769	-0.002	-0.0002	-0.007
	[0.173]	(0.011)	(0.010)	(0.009)
Dryland rice	1.081	0.025	0.032	0.026
	[1.216]	(0.023)	(0.024)	(0.024)
Gram	1.476	-0.004	0.0009	-0.015
	[0.408]	(0.026)	(0.022)	(0.022)
Groundnut	1.404	-0.0003	-0.011	-0.024
	[0.497]	(0.026)	(0.022)	(0.023)
Maize	6.723	-0.036	-0.013	-0.058
	[1.975]	(0.115)	(0.099)	(0.102)
Pearl millet	1.263	0.008	0.024	0.027

continued

Table 6: RD Estimates for Continuity of Covariates

		RD Bandwidth		
Dependent Variable	Mean [SD]	[25 km]	[50 km]	[100 km]
	[1.321]	(0.028)	(0.026)	(0.029)
Pigeon pea	1.914	0.004	0.006	-0.015
	[0.632]	(0.033)	(0.029)	(0.028)
Rapeseed	0.854	0.013	0.005	0.009
	[0.655]	(0.019)	(0.016)	(0.017)
Sorghum	5.917	-0.052	-0.038	-0.045
	[1.35]	(0.100)	(0.086)	(0.083)
Soybean	2.12	0.020	0.027	0.001
	[0.745]	(0.042)	(0.038)	(0.034)
Sugarcane	1.179	-0.005	0.023	0.031
	[1.866]	(0.027)	(0.043)	(0.056)
Sunflower	1.029	-0.004	-0.060*	-0.079*
	[0.724]	(0.023)	(0.035)	(0.044)
Wetland rice	1.727	-0.012	-0.019	-0.023
	[1.091]	(0.037)	(0.037)	(0.046)
Wheat	1.321	0.002	-0.016	-0.026
	[1.141]	(0.034)	(0.032)	(0.032)
Panel D: Social and Demographic Characteristics				
Household size	5.785	0.119***	0.076*	0.024

continued

Table 6: RD Estimates for Continuity of Covariates

		RD Bandwidth		
Dependent Variable	Mean [SD]	[25 km]	[50 km]	[100 km]
	[0.931]	(0.043)	(0.041)	(0.040)
Literacy Rate (percent)	0.503	-0.009	-0.003	0.003
	[0.133]	(0.009)	(0.007)	(0.007)
Log Village Area	6.281	-0.065	-0.051	-0.025
	[1.035]	(0.055)	(0.056)	(0.047)
Log Population	7.455	-0.062	-0.032	-0.040
	[1.056]	(0.050)	(0.044)	(0.038)
Share of Scheduled Caste/Tribe Population	0.283	-0.016	-0.005	0.0005
	[0.228]	(0.018)	(0.014)	(0.012)
Observations	85,745	22,364	44,982	85,745
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. Each cell reports a regression discontinuity (RD) coefficient using regressions as described in Table 3. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 7: RD Estimates for Pollution adjusted for log(fecal coliform)

	RD Bandwidth		
	[25 km]	[50 km]	[100 km]
log(Biological Oxygen Demand)	0.857	0.889*	0.683**
	(0.696)	(0.513)	(0.330)
Observations	1,143	1,844	2,830
R2	0.905	0.847	0.803
log(Chemical Oxygen Demand)	0.689	0.794*	0.551*
	(0.495)	(0.444)	(0.311)
Observations	948	1,526	2,362
R2	0.847	0.776	0.738
log(Electrical Conductivity)	0.513	0.588	0.464*
	(0.380)	(0.400)	(0.257)
Observations	1,109	1,779	2,709
R2	0.941	0.930	0.928
log(Dissolved Oxygen)	-0.169	-0.268	-0.283**
	(0.221)	(0.171)	(0.123)
Observations	1,096	1,776	2,724
R2	0.861	0.774	0.712
Distance	X	X	X
Distance X Downstream	X	X	X
Industry X Year FE	X	X	X

Notes: Robustness estimates for Table 2, adjusting for log Fecal Coliform. *p<0.1; **p<0.05; ***p<0.01. ↩