Industrial Water Pollution and Agricultural Production in India

Nick Hagerty
Anshuman Tiwari *

17 December 2024

Abstract

We study how industrial water pollution affects 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, comparing villages immediately downstream and upstream of each site. To overcome data limitations, we use hydrological modeling to compute spatial relationships and machine learning to predict crop yields from satellite data. We find a large, sudden rise in pollutant concentrations in nearby rivers downstream of sites, but we do not detect lower crop yields on average. Yields do fall in specific areas, but aggregate impacts are small. Likely reasons are that not all farms are exposed, pollution dilutes before reaching crops, and industrial effluent can include beneficial nutrients. Water pollution may have other social costs, but damages to crop yields is probably not one of them.

*Hagerty: Montana State University (email: nicholas.hagerty@montana.edu); Tiwari: Energy Policy Institute at the University of Chicago (email: atiwari2@uchicago.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, Kathryn Baragwanath, Marshall Burke, Esther Duflo, Eyal Frank, Matthew Gordon, Peter Hull, Simon Jäger, Peiley Lau, David Molitor, Alex Oberg, Ben Olken, Sheila Olmstead, Luke Sanford, Molly Sears, and Anant Sudarshan. We thank Gowthami Venkateswaran for sharing the Cost of Cultivation survey data. Anshuman Tiwari gratefully acknowledges financial support from the Grantham Research Foundation and the Environmental Defense Fund.

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). But 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. 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.

Three innovations allow us to relax prior methodological constraints. First, we estimate 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. Second, we use hydrological modeling to precisely determine areas that are upstream and downstream and compute spatial relationships.

Third, we obtain crop yields by predicting them from satellite data using machine learning. No other data source is available at high enough spatial resolution for a spatial regression discontinuity design; even in the United States, aggregate statistics are too coarse and agricultural surveys too sparse. As predictors, 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). We train several models using nationally-representative microdata, then use the best-performing model to generate fitted values for every village in our sample. Our model has nearly four times the predictive power of previous approaches that use individual vegetation indices alone.

Our first main result quantifies 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 surface water pollution at these exact locations, raising omnibus measures of pollution in nearby rivers by three to six times. The amount of water pollution released by these sites has not previously been estimated in publicly available sources.

Our second main result is that crop yields, as predicted from satellite data, are at most only slightly lower in villages immediately downstream of high-polluting industrial sites than in comparable upstream villages in the same year. We estimate a 3 percent decline, but the 95% confidence interval includes 0, and we can reject declines of more than 7 percent, suggesting that even the localized effects of industrial water pollution are small. Since pollution dissipates with distance from its source, effects on crop yields further downstream are almost certainly even smaller.

We do see crop yields fall in specific places where we would expect larger effects. We restrict the sample to three sets of villages likely to be most affected by specific pathways of pollution transport: those served by a canal, those near a river, and those with shallow groundwater tables. We show that these villages have much larger downstream increases in groundwater pollution than the full sample, suggesting greater crop exposure. Their crop yield effects also have larger point estimates, and for the largest we can reject zero: Crop yields fall by 10 percent among villages served by a canal.

Why are the effects small? We find support for three explanations. First, not all crops are exposed to industrial water pollution, even in areas immediately downstream of the source. Although groundwater pollution rises in specific subsets of villages, it does not change much in the full sample. Second, crops are exposed to lower doses of pollution than released at the sites. Industrial sites affect groundwater quality less than surface water quality, consistent with sedimentation, filtration, and radial diffusion reducing pollution concentrations. Third, industrial effluent may have beneficial components than help balance the harms. We find suggestive evidence that sites that release more nutrients have smaller effects on crop yields.

We do not find much evidence for the hypothesis that downstream farmers avert damages through costly input substitution. Effects on agricultural inputs are all zero, except for a small, marginally significant increase in irrigated area that fails to intensify in the more-exposed subsamples. We also do not see follow-on effects on household consumption or poverty rates. Effects on crop quality or human health, either of food consumers or farm workers, are possible (Rai et al., 2019) but beyond the scope of our analysis.

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).

We make two methodological contributions. First, we make progress in applying spatial computation methods to the study of water pollution impacts. We (a) find river locations and compute upstream and downstream relationships among them using only elevation data, (b) construct samples of upstream and downstream data even for point sources not located directly on a major river, and (c) classify villages as upstream or downstream of a point source even when not located on the same river. Our methods have several advantages over the more typical approach of assigning point sources to the nearest point on a river, which can produce inaccurate results for reasons we discuss. Our approach also may help relax data constraints in settings that lack standardized hydrographical data products.[1] In the United States, researchers can rely on the National Hydrography

[1] Garg et al. (2018) also provide useful hydrological modeling methods that are tailored to a slightly different type of research question.

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. In other settings, it can be difficult even to conceptually define upstream and downstream relationships, let alone compute them.

Second, we use machine learning to improve satellite-derived proxies for agricultural production at coarse scales. Our approach bridges a set of papers in economics that use individual vegetation indices as outcomes in causal inference (Asher and Novosad, 2020; Haseeb, 2024) with a vast scientific literature (Weiss et al., 2020; Baylis et al., 2021) that predicts crop yields from remote sensing data using machine learning. The scientific literature generally focuses on a single crop at a time in settings where crop locations are already known. Our challenge instead is to estimate production for all crops across vast areas without data that identifies crops or plot boundaries. We first compare the previous solution of individual vegetation indices to ground-truth crop yield data and find they have low predictive power at the village scale. We then show that machine learning can dramatically improve performance and achieve meaningful predictive power even without crop classification data.

2 Background on Water Pollution and Crop Growth

Manufacturing plants, mines, and other industrial facilities produce a variety of waste chemicals which, if untreated or insufficiently treated, will reach surface or ground water systems. These chemicals include organic compounds (including petroleum hydrocarbons, chlorinated and phenolic compounds, volatile organic compounds, and formaldehydes); heavy metals (including cadmium, lead, copper, mercury, selenium, and chromium); salts and other inorganic compounds and ions; acidity or alkalinity; suspended solids; and oil and grease (Bajpai, 2013; Sudarshan et al., 2023). The particular mix of waste chemicals varies widely and depends on the type of industry; Ahmed et al. (2021) give a detailed breakdown by sector.

Many of these pollutants are toxic in sufficient quantities to animals and plants. Agricultural crops are no exception. Plant growth is known to be 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 depending on the dose, exposure, and the particular mix of pollutants. We can provide a few reference points from controlled 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 aluminum 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. However, it is difficult to know how these effects generalize.

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).

Industrial wastewater can also contain components that are beneficial for crop growth. Effluents from sectors such as food and agricultural processing, and paper and pulp manufacturing contain nitrates, phosphates and potassium—the same chemicals used in fertilizers. Though harmful at excessive concentrations, they can enhance plant growth and yields when applied in appropriate quantities (Hawkins and Risse, 2017; Bedane and Asfaw, 2023; Zhang and Lu, 2024). There is growing interest in using wastewater for irrigation in agriculture, though the focus is more often on domestic wastewater (FAO, 2018). It remains an open empirical question not only how large the impacts of industrial pollution are to crops, but also whether the impacts are negative on net.

2.1 Physical pathways of pollution transport

How does water pollution reach crops? Possible pathways of pollution transport 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. Pollution can reach crops nearly immediately, in the case of surface water irrigation, or accumulate over decades in soil or aquifers. Each of these exposure channels may produce different spatial and temporal patterns of treatment intensity, depending on topography, geology, soils, infrastructure, and irrigation practices.

These exposure channels are neither directly observable nor easy to model. In particular, the behavior of groundwater and its interactions with surface water are highly complex and difficult to model accurately even in data-rich settings. For our main specification, we remain agnostic about the transport pathway. 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 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 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 immediately upstream.

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.

3.1 Hydrological modeling of spatial relationships

We first compute the spatial relationships necessary to construct a dataset of monitoring stations relative to industrial sites. This involves assigning each industrial site to a nearby river and determining where its effluent likely enters the river. We then build on these spatial relationships to construct a dataset of villages relative to industrial sites (for RD analysis of other outcomes).

We use hydrological modeling to compute these relationships accurately. Instead of relying on an existing map, we use elevation raster data to model where runoff flows; rivers emerge where streamflow accumulates. Using this model, we can calculate the flow line (i.e., route) that would naturally be taken by water released at any point on a map. We then calculate relationships by comparing flow lines.

This type of modeling is routine in water resources and related fields; it is highly accurate at predicting the locations of rivers. It relies only on basic tools available in ArcGIS Pro for which tutorials are widely available, so it can be used in other studies that need to accurately characterize relationships on surface water networks.

Defining the river for each pollution source. Our approach is illustrated in Figure 2. This figure shows our research design for one site in our sample: Jharsuguda, a major industrial hub in the state of Odisha. The industrial site is represented by the orange dot.

We use our hydrological model to construct what we call a "reference" flow line, shown in blue, for each industrial site. The reference flow line is a continuous streamflow path (i.e., from source to ocean) satisfying three criteria: (1) it receives natural drainage from the industrial site, (2) the point at which the drainage enters the river is relatively close to the site itself, and (3) it extends upstream as far as possible into areas unaffected by the site. We construct this path by tracing the industrial site's own flow line to a point 25 km downstream and then following flow lines both upstream and downstream of that point. We detail these methods (and make the criteria precise) in Online Appendix 1.1. Our sample of monitoring sites is then comprised of those that fall along each industrial site's reference flow line.

Defining the treatment variable. Defining whether a monitoring station is downstream or upstream of the pollution source is equivalent to determining the point at which industrial pollution enters the river. Since this point is unobserved—effluent may follow a canal or ditch instead of its natural flow line—we consider several candidate definitions and test them empirically. The best-performing definition is based on flow length. Flow length is the total length of the flow line from a given point to the ocean; it measures how far upstream a point is located within a watershed.

We therefore classify a monitoring station as downstream of an industrial site if it has a shorter flow length than the site, and upstream otherwise. This treatment definition is well-grounded in basic physics: It allows for effluent to move diagonally across contour lines (e.g., via a ditch or through groundwater), but not upstream, against the gradient of flow.[2]

The alternative treatment definitions we considered are based on: (1) position relative to the intersection point between the reference flow line and the industrial site's flow line, i.e., the point where drainage would naturally enter the river; (2) position relative to the nearest point on the river, as in most prior literature; and (3) elevation relative to the industrial site, with lower elevation classified as downstream (Asher et al., 2022). We tested them by comparing their RD estimates of industrial sites on surface water pollution concentrations. The treatment variable based on flow length produced the strongest "first stage" effects, while others were smaller and often not statisti-

[2]Let us be more precise (it may help to refer to Figure 2). For monitoring stations far downstream or far upstream of the industrial site, it is clear whether they receive effluent from the industrial site, and flow length classifies them correctly. For example, consider a monitoring station downstream of the point of intersection between the reference flow line and the industrial site's flow line. Its flow line fully coincides with part of the industrial site's flow line, but its flow length is shorter, so it is classified as downstream. For a monitoring station close to the industrial site but not on the site's flow line, it may or may not receive effluent. But empirically, comparing their flow length appears to do well at classifying them correctly.

cally significant.[3]

Constructing the village sample. For the village-level dataset, we include all villages that fall within 20 km of the reference flow line. This span gives us plenty of data to work with while focusing analysis on areas most likely to be affected by pollution. We maintain the same definition of the treatment variable as for monitoring stations, classifying a village as downstream if it has a shorter flow length than the industrial site, and upstream otherwise. The resulting sample is shown in Figure 2, with downstream villages in light green and upstream villages in dark green.

This approach captures the essential intuition of comparing “downstream” and “upstream” villages despite the fact that these terms lack clear meaning when applied to villages instead of river segments. To define which villages are downstream of a pollution source, we need to make assumptions about which villages are potentially affected by pollution. Pollution can be transported away from its flow line via multiple possible mechanisms and we want to capture all of them. Using a moderate radius for sample selection focuses analysis on the areas closest to the surface route that pollution would naturally travel, while allowing for the potential for pollution to affect the surrounding areas. And because our upstream villages are selected through the same criteria as downstream villages (i.e., using the reference flow line), we avoid introducing mechanical discontinuities that can result from asymmetric selection criteria.

Advantages over prior work. Researchers using similar designs often simply “snap” the pollution source points to the nearest point on the nearest river encoded in a published shapefile (e.g., He et al. (2020)). But this method can introduce potentially severe measurement error and other problems if pollution sources are not located immediately adjacent to a major river. Such locations are not rare. For example, consider the industrial sites in our sample, mapped in Figure 3 against a coarse shapefile of major rivers. Many sites that appear to be located near major rivers are in fact several kilometers away. Other sites are located far away from any of the rivers shown on the map.

In such cases, four specific problems can arise. First, if the river network is coarse, the source can be snapped to a river far away, missing closer areas of greatest exposure. Second, if the river network is detailed, the source can be assigned to a small stream that does not extend very far enough upstream, leaving insufficient data for a control group. Third, the nearest point may not be where pollution actually enters the river, resulting in false downstream and upstream classifications. Fourth, the nearest river may not even receive the effluent at all. For example, one industrial site

[3] Another way of thinking about flow length is that it is a kind of compromise between two bounds. Effluent can follow its natural flow line, or potentially move diagonally across contour lines, but it cannot flow uphill. So effluent must enter the river below the elevation of the industrial site, and it probably enters the river at its natural flow line if it hasn’t already. Generally, the point of equal flow length falls somewhere between the point of equal elevation and the point of intersection with the site’s flow line.

in our sample 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.

Hydrological modeling allows us to avoid these problems. The reference flow line controls the coarseness of the river network, guaranteeing a nearby river with as much upstream data as possible. The model generates multiple potential definitions of where pollution enters the river, which can be tested with water quality data. And tracing the industrial site's flow line ensures correct identification of the rivers that receive industrial effluent.

Problems with the snapping method are exacerbated in village-level analysis, since it can misclassify the treatment variable for villages close to the pollution source. For example, in Figure 2, a number of villages to the immediate southeast of the industrial site would be classified as downstream based on their nearest point on the river, even though they have longer flow lengths. Our approach instead gives precise treatment classifications for villages arbitrarily close to the industrial site. This precision is crucial for an RD design in which we expect the greatest effects to be closest to the site itself, potentially before reaching the river.

3.2 Geographic regression discontinuity

Our main analyses estimate the local effects of being immediately downstream of a heavily-polluting industrial site. We set up a multi-cutoff geographic RD following Cattaneo et al. (2024). We pool data across industrial sites, normalize by distance to the site, and estimate the mean difference in outcomes approaching a site from downstream versus from upstream:

$$\tau = \underbrace{\lim_{Distance \downarrow 0} \mathbb{E}[y_{ist}|Distance_{is} = 0]}_{\text{Downstream}} - \underbrace{\lim_{Distance \uparrow 0} \mathbb{E}[y_{ist}|Distance_{is} = 0]}_{\text{Upstream}} \tag{1}$$

in a dataset consisting of the villages (or monitoring stations) i that belong to the sample for each industrial site s, across all observed years t. The score (i.e., running variable) $Distance_{is}$ is the geographical distance between the observation i and its site s.[4] Its sign is set to positive for downstream villages and negative for upstream villages.

We estimate Equation 1 via local linear regression on each side of the cutoff without higher order polynomials (Gelman and Imbens, 2014) and with a triangular kernel (Fan and Gijbels, 1996). We select separate bandwidths for each outcome using the optimal bandwidth algorithm of Calonico et al. (2020). We adjust for site-by-year fixed effects to improve precision and avoid bias from

[4]We use geographical distance as the score because we want to estimate local effects at the point of the site. Alternatives such as river distance or flow length would set up a boundary discontinuity, estimating the difference in conditional expectations at all points along the line of (e.g.) equal flow length rather than at the site itself. Another way of putting this is that kernel weights decline radially with geographical distance but bilaterally with the alternatives.

differential balance across sites. They ensure our effects are estimated using only the variation between upstream and downstream observations for the same industrial site in the same year. Therefore, in practice, our estimates come from regressions of the form:

$$y_{ist} = \tau Downstream_{is} + \gamma Distance_{is} + \delta Distance_{is} \times Downstream_{is} + \alpha_{st} + \varepsilon_{ist}. \quad (2)$$

We report the robust confidence intervals and p-values of Cattaneo et al. (2024). We cluster standard errors by subdistrict, the administrative division above village, to account for correlation across space and time in both bandwidth selection and inference. Clustering also accounts for repeated observations, when the same village appears more than once in the pooled sample for different industrial sites.

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 had been strategically placed 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 more likely to shift to urban areas rather than the rural areas immediately upstream because of India's rigid land and labor markets (Hsieh and Klenow, 2009; Duranton et al., 2015).

3.3 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 it moves downstream—pollutants can break down, deposit on streambeds, or become diluted as a river collects runoff and joins other tributaries. Impulse response functions let us extrapolate more formally. We describe the estimation procedure in Online Appendix 1.2.

3.4 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 unobserved, 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.

In addition to these conceptual disadvantages, temporal variation is 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. Although most of these sites have grown over time, this growth is correlated across sites over time as India has industrialized, leaving little useful variation, and available measures of industrial plant growth are noisy.

4 Predicting Yields Using Satellite Data

To generate agricultural outcome data at a high spatial resolution, we derive a measure of crop yields from satellite data. We use machine learning to extract the most information possible from satellite data. We train a predictive model using a sample of village-level ground-truth data, and then we use the model to predict crop yields for every village in our analysis sample. Our model generalizes prior approaches that proxy for crop yields using individual satellite-derived indices (e.g., Asher and Novosad (2020)), as well as those that combine multiple indices using linear regression (e.g., Lobell et al. (2020)). Our objective is not to perfectly predict crop yields but rather to improve upon these previous approaches for use in causal inference.

4.1 Data

Data sources and processing are summarized here and detailed in Online Appendix 1.3.

Vegetation indices. The remote sensing literature has proposed a number of measures to proxy for crop yields, called vegetation indices (VIs). We use six 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). To this list we add the Enhanced Vegetation Index (EVI) used by Asher and Novosad (2020) and Asher et al. (2022).

Predictors. We extract minimum and maximum values of each VI and their underlying bands during agricultural years 2015-17 from the Sentinel-2 MSI satellite and aggregate them to villages.

Village-level crop yields. To train our model, we use plot-level microdata from the Indian government's Cost of Cultivation survey. For each village, we calculate crop yields per hectare for agricultural years 2015-17 and average across sampled plots, weighting by crop prices and plot area. We refer to the outcome variable as the revenue value of yield and predict its log-transformed value.

4.2 Model training

Building the predictive model consists of four steps: (1) tune each candidate model, (2) select the best model, (3) evaluate its performance, and (4) generate predictions. We randomly split the Cost of Cultivation data ($n = 1793$) into three distinct sets: a 64% training set (for tuning), a 16% test set (for model selection), and a 20% evaluation set (for measuring performance). We use distinct test and evaluation sets because the model selection step is itself part of the model training process and can be overfit.

We tune three models: elastic net, random forests, and boosted trees. Elastic net is a type of regularized linear regression that nests both ridge and lasso regression. Random forests and boosted trees are nonlinear models that create ensembles of decision trees, which recursively partition one variable at a time. Further details are in Online Appendix 1.3.

4.3 Model selection

Table 1, Panel A summarizes the performance of each candidate model in the test set. Of the three main models we tune, the random forest does the best (greatest R-squared and smallest RMSE). It explains more than twice the variance in crop yields as the elastic net model does, and the boosted trees model is not far behind. The fact that the nonlinear models do so much better than the linear model suggests that the true relationship between reflectance and crop yields is highly nonlinear.

We also find it is important to include both the VIs and raw bands as predictors. Rows 5 and 6 show that random forest models trained on only VIs and on only the raw bands perform similarly to each other, and considerably worse than the model that uses both. This result suggests that the VIs provide structure that is valuable when ground-truth training data is limited, but the bands also contain useful information that is not fully captured by the functional form of the VIs.

4.4 Evaluating alternative proxies

Besides our machine learning models, we evaluate two alternatives for proxying for crop yields (Table 1, Panel B). The first is the common approach of simply using individual VIs directly. Following Asher and Novosad (2020), we regress our observed log crop yields on the log of the difference between maximum and minimum values of NDVI. The regression coefficient is positive and statistically significant, but its out-of-sample predictions are poor. Its test-set R^2 is low, only about one-quarter as large as that of our best machine learning model.

The second is to fit a model using district-level data, and then use it to make village-level predictions. The advantage is that district-level data covers all of India and is based on a much larger sample than the Cost of Cultivation survey. The disadvantage is that the structure of the relationship between reflectance and yields may vary at different spatial scales, so the model may not downscale well. Online Appendix 1.4 describes the details of our district-level predictive model. Satellite data predicts district-level crop yields better than village-level yields, but the district-level model does poorly at village-level prediction. A linear regression with all our VI predictors produces an in-sample R^2 of 0.39 (Online Appendix Table 2), well exceeding our best village-level model. However, its out-of-sample performance at the village scale is less than 0.05 (Table 1).

4.5 Model performance

We now use the evaluation set to estimate the out-of-sample performance of the random forest model. Online Appendix Figure 1, Panel A plots predicted values against observed values in the evaluation set, and Table 1, Panel C shows that the model achieves an R^2 of 0.25.

We interpret this as good performance for two reasons. One, we are trying to predict yields across all crops for an entire nation. Since different crops look very different in satellite images, this is an inherently much noisier task than typical yield prediction projects, which tend to focus on a single crop. For example, Lobell et al. (2020) report an R^2 of 0.58 in data that includes exact plot boundaries for a homogeneous crop (maize) in a small geographical region. In contrast, our data is spread across a much larger region and includes all crops and land uses in the country. This context makes our model's performance more impressive.[5]

Two, we are evaluating our model against noisy estimates of our prediction target, not its true values. We want to predict average yields for the whole village, but our training data comes from only a small sample of plots in each village. The population means of all plots in each village would have lower variance than these sample means, so our predictions would likely explain a

[5] Our model would of course be improved by incorporating high-resolution crop identification data, but such data do not yet exist for India. 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 even after extensive cleaning we found the data quality too low to be useful.

much greater share of the variance in population means. In other words, the R^2 in this sample is an underestimate of the R^2 for entire villages.

4.6 Generating predictions

We use the random forest model to generate our main outcome variable, predicted log yield, for all villages in our RD analysis sample for the year 2015. Online Appendix Figure 1, Panel B plots the distribution of predicted values in the analysis sample on top of the distribution of observed values in the Cost of Cultivation data. As usual, the predicted values have lower variance. Otherwise, the model predictions do not fall outside of values seen in the training data, which suggests that the analysis sample is fully within the support of the training sample. This helps to reassure us that the predictions are reliable.

5 Data and Summary Statistics

5.1 Other Data

Industrial sites. India's Central Pollution Control Board (CPCB) selected 88 industrial sites for detailed, long-term study in 2009. Names of these sites are taken from the CPCB document "Comprehensive Environmental Assessment of Industrial Clusters" (Central Pollution Control Board, 2009). We identify the geolocation of each site using Google Maps and other publicly available reference information. These sites are displayed as orange dots in Figure 3.

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.

Surface water quality. We use 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 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 manually verified, refined, or corrected the geolocation of each station by 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 3.

Many water quality parameters have been collected by the CPCB at some point. However, only a few parameters are measured consistently. We focus on four common omnibus measures that proxy for a wide range of pollutants: chemical oxygen demand (COD), biochemical oxygen demand (BOD), dissolved oxygen saturation (DO), and electrical conductivity (EC). 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). EC is used to measure salinity or inorganic compounds, since the ions created by dissolved salts and minerals are what allow water to conduct electricity. We also show results for a number of less consistently reported parameters. Finally, we calculate an indicator for whether the water meets the CPCB's Class E surface water criteria, for irrigation, industrial cooling, and controlled waste disposal.[6]

Groundwater quality. We also gather measurements of groundwater pollution, collected by several central and state government agencies and made available through the India Water Resources Information System (IndiaWRIS) portal. The data include biannual observations from $14,704$ monitoring stations throughout the country between 2000 and 2022, including location coordinates. We geolocate monitoring stations within villages and construct annual village-level means of available water quality parameters. To minimize the influence of reporting errors and other extreme values, we winsorize each parameter at its 95th percentile.

Again only a few parameters are measured consistently, and they are different from the parameters most frequent in the surface water quality data. COD, BOD, and DO are unavailable, so we focus on four other measures. Two are EC and total dissolved solids (TDS), which measures the total amount of inorganic and organic material in the water. For the third, we create a "high pollution indicator" for whether any available parameter exceeds its 90th percentile. The groundwater data include few omnibus measures but many specific ones, so this indicator is a way of incorporating all the parameters available while reducing their dimensionality. Fourth, we calculate an indicator whether the water meets the CPCB's Class E groundwater criteria, for industrial and controlled

[6]These criteria are: pH between 6.0 to 8.5, EC below 2250 μmhos/cm, sodium absorption ratio below 26, and boron below 2 mg/L (https://indiawris.gov.in/wris/#/SWQuality). Boron measurements are not available in the data, so we calculate the indicator based on the first three criteria.

waste disposal.[7]

Village covariates. We use the Population Census of 2001 for baseline village covariates, the Population Census of 2011 for agricultural inputs and village outcomes, the Economic Census of 2013 for employment in polluting industries, potential yields from GAEZ for cropland quality and crop suitability, village boundaries from NASA's SEDAC, and harmonized village definitions from SHRUG. Details are provided in Online Appendix 1.5.

For indicators of irrigation source availability, we obtain a national geospatial dataset of canal lines from IndiaWRIS. We classify a village as served by a canal if any canal from this shapefile intersects the village's boundaries. We measure distance to river for each village by calculating the geographic distance between its centroid to the flow line of the corresponding industrial site. We calculate depth to groundwater by inverse distance kriging (i.e., weighted spatial interpolation) across pre-monsoon measurements downloaded from IndiaWRIS, taking means within wells across years 2014-16 (to match our yield measurements) and means within villages across raster cells.

5.2 Continuity tests and summary statistics

We provide summary statistics in Table 2 for our main outcome variables on pollution and agricultural output.

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 estimating the geographic RD parameter from 1 with each covariate on the left-hand side. For the RD design to be valid, 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: (a) physical characteristics, (b) potential yields estimated for common crops, (c) commercial and public amenities, and (d) social and 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.

[7] These criteria are: TDS less than 2000 mg/L, sodium absorption ratio less than 18, and pH between 6.0 to 8.5 (https://indiawris.gov.in/wris/#/GWQuality).

Figure 4 shows visual evidence of continuity for a selection of these covariates. For context, we first plot a histogram of village observations. The density falls symmetrically near the industrial site because we are conducting a geographic RD at a single point—the area that falls within a given radius increases linearly with that radius, until the sample width becomes constrained by the 20-km buffer around the reference flow line. 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.

In the rest of Figure 4, all other variables appear to be continuous. Plots of potential yields for each specific crop are in Online Appendix Figure 4; they also appear continuous. Confidence intervals and RD estimates for these covariates and many others are shown in Online Appendix Table 1. Across the 31 variables we test, only one is statistically significant at a 5% or even 10% level: whether a banking facility is available in the village. Since we lack a mechanism to explain this apparent discontinuity, we attribute it to expected sampling variation.

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 are affected by being downstream of these industrial sites. In robustness checks, we control for all these covariates.

6 Effects on Pollution

6.1 Surface water

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 five key water quality measures: chemical oxygen demand (COD), biological oxygen demand (BOD), dissolved oxygen (DO), electrical conductivity (EC), and violation of the CPCB's Class E criteria. The graphs plot mean values of each measure 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. We also fit fourth-order polynomials to show global patterns.

All five measures show a discontinuous increase in pollution at the exact location of the industrial sites. COD, BOD, EC, and Class E violations increase; these measures are undesirable, with higher levels indicating worse water quality. DO decreases, which also indicates an increase in pollution; this measure is desirable, with lower levels indicating worse water quality.

Table 3 quantifies these results. It reports the geographic RD parameter from Equation 1, estimated as described in Section 3.2, separately for each water quality measure. Each estimate represents the increase in the dependent variable immediately downstream of an industrial site, adjusting for site-by-year fixed effects.

The estimates are quantitatively large. For example, the estimate of 68.3 for COD implies that the average "severely polluted" industrial site more than triples pollution levels in nearby rivers relative to the sample mean. Confidence intervals easily exclude zero at a 95% level for all five measures.

Online Appendix Table 4 reports RD results for 16 additional water pollutants available in CPCB data. Nearly every reported pollutant worsens by a large and statistically significant amount. This is true for measures of salinity (presence of ions like calcium, chloride, magnesium, and sodium), nutrients (nitrates, nitrites, potassium, and sulphates), acidity or alkalinity (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.

Moving beyond local effects, 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 measure, showing how industrial clusters affect river pollution over the course of the river. For all five measures, the increase in pollution is greatest immediately after the industrial site. It then steadily falls and rejoins the trend implied by the upstream curve around 100 km from the industrial site. Dissipation could result from several processes: sedimentation, chemical or biological degradation, wider diffusion into aquifers, and/or dilution by entering tributaries.

6.2 Crop exposure and transport pathways

Industrial sites release a lot of pollution. Does this pollution actually reach crops, and if so, how?

To answer these questions, we study groundwater quality. Since we cannot directly observe crop exposure to water pollution, groundwater quality is the best alternative. It is a useful proxy for two reasons. First, many crops are irrigated with groundwater. For them, measurements of groundwater quality *are* nearly direct measurements of pollution exposure. Second, groundwater collects pollution from all sources of irrigation water, since a fraction of applied water percolates down into the aquifer. If polluted water from rivers or canals is used for irrigation, then the pollution is likely to be reflected in the groundwater quality.

We first estimate downstream effects of the industrial sites on groundwater quality in the full sample (Table 4, Panel A). Overall, the sites have little effect on pollution in groundwater, in contrast to the effects in surface water. The one parameter we can directly compare between groundwater and surface water is EC. Its estimate is statistically significant, indicating an increase in salinity, but its magnitude is only six percent as large as in surface water. Estimates for other measures—total dissolved solids, an indicator for high pollution in any reported parameter, and Class E violations—are small and insignificant.

Next, we test whether industrial pollution reaches crops through the specific pathways of pollution transport described in Section 2.1. We do so by restricting the sample, both upstream and downstream, to villages most likely to be affected by each specific pathway. Even though effects in the full sample are small, it is possible that they are hiding meaningful heterogeneity.

We investigate the three main potential pathways: canals (via irrigation or percolation from unlined canals), rivers (again via irrigation or percolation), and groundwater diffusion (directly through the aquifer). We restrict the sample using fixed physical characteristics that are unlikely to endogenously respond to downstream pollution, avoiding the worst forms of selection bias. For canals, we restrict the sample to villages that any canal passes through, using our geospatial dataset of canal lines. For rivers, we restrict the sample to villages whose centroid falls within 5 km of the reference flow line.

For groundwater diffusion, we restrict the sample to villages with shallow water tables, for two reasons. First, areas with high water tables are more likely to be closely connected with both surface water systems (so they can easily receive the pollution) and each other (so they can transmit it). Deep aquifers are more likely to be separated both vertically and horizontally by rock or sediment with low permeability. Second, areas with shallower water tables have lower pumping costs, since less energy is needed to move the water to the surface. We use a maximum depth of 8 meters, the threshold at which centrifugal pumps no longer function and more expensive submersible pumps must be used (Sekhri, 2014).[8] Each subsample is a relatively small fraction of the full sample (between 8 and 19 percent).

We find evidence for pollution transport through all three pathways. Table 4, Panels B-D show that industrial sites affect downstream groundwater quality in all three subsamples. Evidence is strongest for villages close to the river (Panel C). All four measures increase downstream and have confidence intervals that exclude zero. Still, the effects are smaller than for surface water—EC increases in groundwater by only 38 percent as much as it does in the river itself.

[8]An endogenous response in this variable is possible, but only for a small subset of villages near the threshold, so it would be unlikely to change the overall results.

Evidence is also solid for the other two pathways. For villages served by canals (Panel B), TDS and the high pollution indicator both increase downstream by a large amount, though not all effects on water quality are bad: EC goes down, indicating a decrease in salinity. For villages with a shallow water table (Panel D), the high pollution indicator increases by a large share, and EC and TDS also increase.

Overall, the evidence from groundwater quality supports three conclusions. First, water pollution from industrial sites does reach crops, and it likely does so through multiple pathways. Second, this pollution has a relatively limited reach: it affects only a subset of villages most directly exposed to these transport pathways. Third, the level of pollution that reaches crops is lower than measured in nearby rivers—perhaps because of radial dilution, sedimentation and filtration, and/or higher flow rates in rivers.

7 Effects on Agriculture

7.1 Crop yields

Having shown that industrial sites increase pollution, we investigate how this pollution affects agricultural production in downstream villages, using our measure of crop yields predicted from satellite data. We first report results for the full RD sample, and then for the subsamples of villages most affected by specific physical pathways of pollution transport.

Full sample. Figure 6 visualizes our main result for crop yields. It shows RD plots similar to those for pollution, but for the predicted log revenue value of yield. The first plot uses raw data; the second adjusts for industrial site fixed effects.[9] The plots hint at a discontinuous drop in crop yields at the industrial site, but any such drop is small and not obviously distinguishable from background variation. Informally, if we visually extrapolate away from the RD threshold of 0, any impact of pollution appears to quickly dissipate as crop yields rejoin the broader trendline within 25 to 50 km.[10] Despite increasing surface water pollution drastically, industrial sites do not seem to have a major effect on downstream crop yields.

Table 5 quantifies this result. The RD estimate for predicted log crop yield in the base specification (Panel A) implies that crop yields are 3 percent lower immediately downstream of a severely-polluting industrial site. However, this effect is not statistically different from zero. The 95%

[9] For RD plots without fixed effects, we use the `rdrobust` package in R, with IMSE-optimal quantile-spaced bins. For RD plots with fixed effects, we use bin definitions and global polynomial fits from `rdrobust`. Since this package is unable to adjust binned means for covariates, we use `binsreg` to calculate covariate-adjusted binned scatter points, evaluating both global polynomials and binned points at the mean of the fixed effects.

[10] In the plot with fixed effects, there also appears to be a small, symmetric dip in crop yields on both sides of the RD threshold; this is likely driven by farmland conversion and error in the cropland mask close to the industrial sites.

confidence interval allows us to reject reductions in crop yields larger than about 7 percent. Although a 7-percent or even 3-percent effect on aggregate crop yields would perhaps constitute a severe impact to production, recall our RD design estimates a local treatment effect for the villages most directly affected by industrial pollution. Since pollution rapidly dissipates away from the sites, we can expect the impacts further downstream to be much smaller.

We report robustness checks in Panels B-D of Table 5. 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 Online Appendix Table 1. Panel D controls for irrigation-related agricultural input variables listed in Table 5.[11] All these specifications produce similar results as the main specification. None of the estimates are statistically different from zero.

Villages exposed to specific pathways. Small or zero effects in the full sample are consistent with our analysis of groundwater quality, which suggests that industrial pollution does not reach crops in most downstream villages in high concentrations. But we do see evidence that industrial pollution reaches crops in certain villages that are affected by specific pathways of pollution transport. Next, we ask whether crop yield effects are stronger in these villages.

Results in the first panel of Table 7 and in Online Appendix Figure 2 offer a tentative yes. Point estimates for all three subsamples are larger than the estimate for the overall sample. For villages served by a canal, industrial sites reduce crop yields immediately downstream by 10 percent, and we can reject a null hypothesis of no effect. Estimates for near-river and shallow-groundwater villages are also larger than for the overall sample, though we lose precision with less data, so confidence intervals still include zero. Taken together, this evidence suggests that crop impacts are greatest in the places we would expect them to be greatest.

Why are pollution impacts worst for villages with canals? Canal irrigation may provide the most direct exposure to industrial effluent. Other pathways are likely to involve at least some transport through aquifers, which can filter some of the pollutants.[12] Although groundwater quality is not clearly worse in canal villages than the other two subsamples, the water applied to crops may very well have higher pollution concentrations than the groundwater. It is also possible that the specific types of pollutants that reach villages through canals (rather than being filtered) are worse for crops.

[11] We omit robustness checks that vary the RD bandwidth, since Cattaneo, Idrobo, and Titiunik (2020) argue they are inappropriate. Bandwidths that are much larger or smaller than optimal will introduce too much bias or variance, making point estimates unreliable and invalidating the robustness check itself.

[12] Even for villages near rivers, relatively little cropland is reported as irrigated directly from the river. Instead, pollution more likely reaches crops by traveling through the river and then the aquifer.

7.2 Agricultural inputs and household welfare

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 6 reports RD estimates for a set of agricultural inputs in the full sample, and plots are provided in Online Appendix Figure 3. Neither land (as measured by crop area as a share of village area) nor labor (share of employment in agriculture) change. Irrigation, probably the most obvious margin of adjustment, may expand. We estimate that the share of crop area under irrigation increases by 6 percentage points, though the evidence for a positive effect is not strong ($p < 0.07$). One might expect farmers to avoid irrigation if the water is harmful to crops, but if water quantity can substitute for quality, farmers might instead irrigate more to compensate for the harm. However, we find no evidence that farmers substitute between irrigation sources—effects on the share of irrigation from canals, wells, tanks or lakes, and other sources including rivers are all small and insignificant.

Finally, we test for effects of industrial pollution on household welfare, as measured by per-capita consumption and the poverty rate. We find no effects for either.

Villages exposed to specific pathways. We again look at the subsamples in which groundwater pollution and crop yield effects are largest. Table 7 reports results. All estimates are small and statistically insignificant. The point estimates for irrigated area are smaller than in the full sample, but their differences are not significant. We also see no effects on specific irrigation sources in the subsamples we might expect: canal irrigation in villages served by canals, and well irrigation in villages with shallow water tables. Overall, we do not find much evidence that a small effect on crop yields is an equilibrium result of input adjustment.

7.3 Does some industrial effluent benefit crops?

Industrial effluent often includes salinity, heavy metals, and toxic organic compounds that are known to harm crops. But it can also include nitrates, phosphates, and potassium, which are the components of fertilizer and can benefit plants as nutrients. These potentially beneficial pollutants can also be found in domestic and municipal effluent (i.e., untreated sewage), which is released by the towns and cities that often coincide with industrial sites (National Academies, 1996; Hussain et al., 2002; Abdoli, 2022). Perhaps the effluent from industrial sites contains beneficial nutrients in addition to harmful pollutants, and they partially offset each other, leading to small net effects.

We attempt to test this hypothesis by estimating effects of different groupings of industrial sites on crop yields. In Online Appendix 1.6, we find suggestive evidence that crop yield effects are concentrated among sites expected to have more industrial effluent relative to municipal effluent, and among sites that release relatively little nitrate. Although none of the estimates is precise, we view the available data as lending very tentative support to the beneficial-nutrient hypothesis.

8 Discussion

8.1 Contextualizing the results

Our results suggest that the aggregate real-world harms to crop yields from industrial water pollution are small. Some villages experience damages, particularly those served by canals. But on average, we can reject declines in crop yields of more than 7 percent in villages immediately downstream of industrial sites. And this is a local effect—since we show pollution dissipates further away from the sites, it represents an upper bound for the overall impacts of industrial sites on crops. Damages of 3 or even 7 percent would indeed be harmful to farmers in the affected area, but this upper bound would apply only to a very small region. Assuming crop yield impacts scale with pollution concentrations, crops more than 50 to 100 km downstream of the sites would be essentially unaffected. Our study also focuses on the most highly polluting industrial sites in India, so the effects of other pollution sources should be smaller.

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., 2023), 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). Plus, all these shocks affect large swaths of the country, not just a small radius around a handful of sites.

8.2 The potential role of measurement error

Even though our proxy for crop yields improves upon previous approaches that use satellite measures, substantial measurement error likely remains, and it may affect our RD estimates. Unfortunately, neither the magnitude nor the direction of bias is clear. Remote sensing measures such as ours, especially those created from machine-learning methods, are known to have non-classical measurement error (Alix-García and Millimet, 2023), so the estimate is not necessarily attenuated toward zero. We do take several precautions to try to minimize measurement error, such as applying cloud and cropland masks. In particular, we spatially aggregate data by village instead of using pixel values directly, a procedure that Garcia and Heilmayr (2022) show can help to reduce bias from measurement error in satellite data. Finally, we are reassured by Proctor et al. (2023), who find that bias in parameter estimates is relatively low when the satellite measure is the outcome variable, as it is in our study, rather than the treatment variable.

Measurement error in traditional survey measures of crop yields is also high and non-classical (Kosmowski et al., 2021); Lobell et al. (2020) show that satellite measures can perform better. In addition, the best available ground-based data is much coarser. We attempt to adapt our main analysis to ICRISAT district-level data in Online Appendix Table 5. As expected, estimates are too imprecise to be useful.

More generally, we note that satellite-derived measures have enjoyed widespread success in the economics and scientific literatures as 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 using a similar RD design. There is strong reason to believe vegetation indices are well-suited to pick up the specific negative impacts of industrial water pollution on crops: Many of the agronomy studies on water pollution in controlled settings report negative impacts to leaf size and color, characteristics that vegetation indices are specifically tailored to measure.[13] Many questions and uncertainties remain about the capabilities of satellite data in applications like ours, but we leave their resolution to future work.

8.3 Explaining the small effects

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 very much. Our analysis uncovers three leading reasons why. First, not all crops are exposed to industrial water pollution, even in areas immediately downstream of the source. We show that water pollution from industrial sites does reach crops, but only along the route of specific transport pathways. Second, pollution is diluted before it is taken up by crops. We show that pollution concentrations are lower in local groundwater than in nearby rivers, likely due to sedimentation, filtration, and radial diffusion. Third, industrial water pollution has beneficial components that may help balance the harms. We find suggestive evidence that sites that release more nitrates, or that have more municipal effluent relative to industrial effluent, affect crop yields by less than others.

[13] One margin of adaptation our analysis may miss is if 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 high-resolution crop classification datasets are not yet available.

Three additional possibilities are worth mentioning. First, it remains possible that farmers adjust agricultural inputs to avert pollution damage. We estimate that irrigation increases downstream, although this effect fails to intensify in subsamples as expected. We find null effects for many types of inputs, though other margins of adjustment remain unobserved. Second, pollution might harm output 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. Crop prices might allow us to measure some (likely not all) quality effects, but such data are not available at high spatial resolution. Third, the previous literature may exhibit publication bias. The case studies that show large impacts of industrial water pollution on crops might be unrepresentative of the true overall effects of pollution.

9 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 predict 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 not detectably lower in villages immediately downstream of these sites on average. They do fall by up to 10 percent in certain subsets of villages that receive more pollution, but this effect likely dissipates rapidly downstream. Third, we show that crop yield effects are likely small because (a) industrial water pollution does not actually reach most crops; (b) when it does, it is in lower doses than seen in nearby rivers; and (c) it contains not only toxic chemicals but also nutrients that act as fertilizer.

Our results do not imply that industrial water pollution is not costly to society, only that agriculture may not be the locus of those costs. There are many other types of potential social costs that we do not quantify, including harm to human health and to ecosystems. We leave these as important objects of future research.

Figures

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

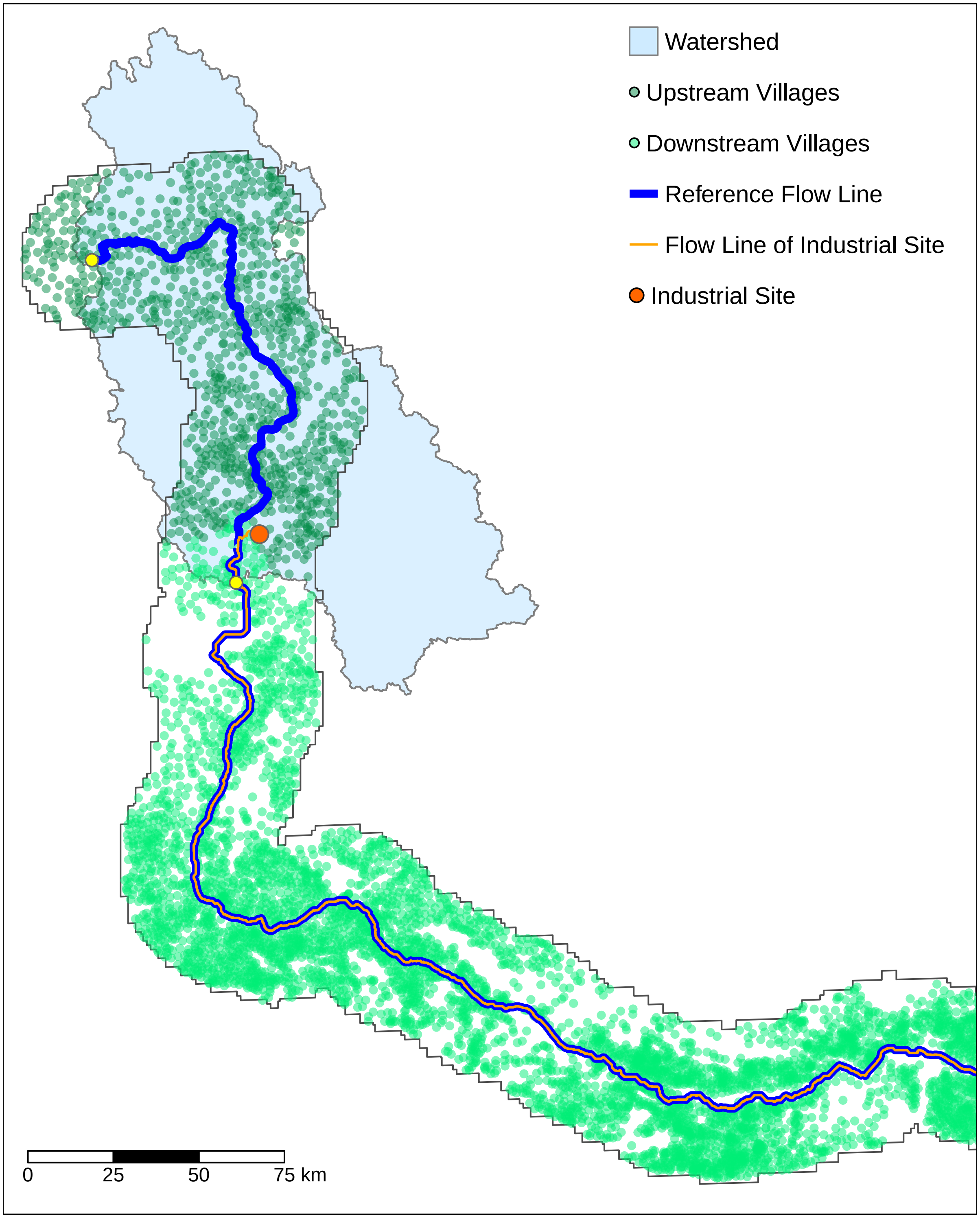

Figure 2: Illustration of the sample selection and treatment assignment for our research design. The site shown is Jharsuguda, a metallurgical hub in the state of Odisha.

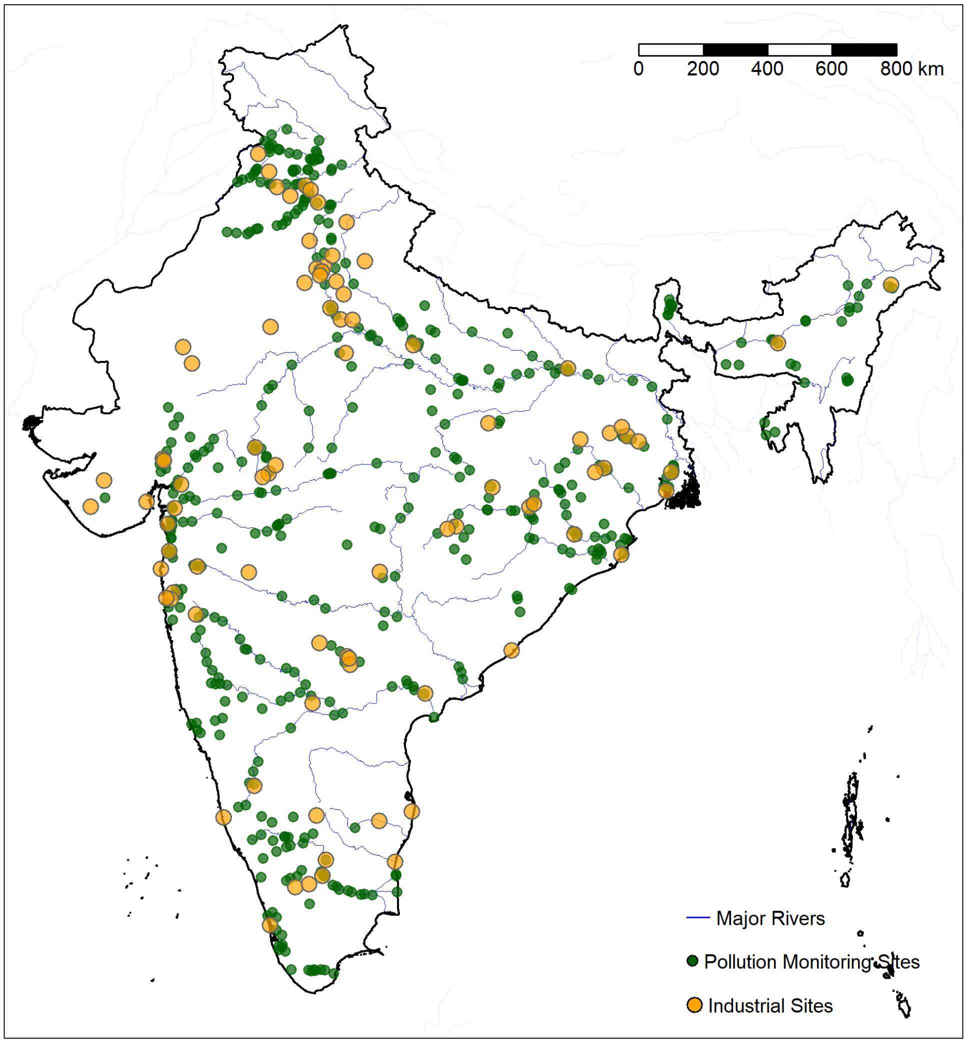

Figure 3: Locations of "severely polluted" industrial sites (orange dots) and water pollution measurement stations (green dots).

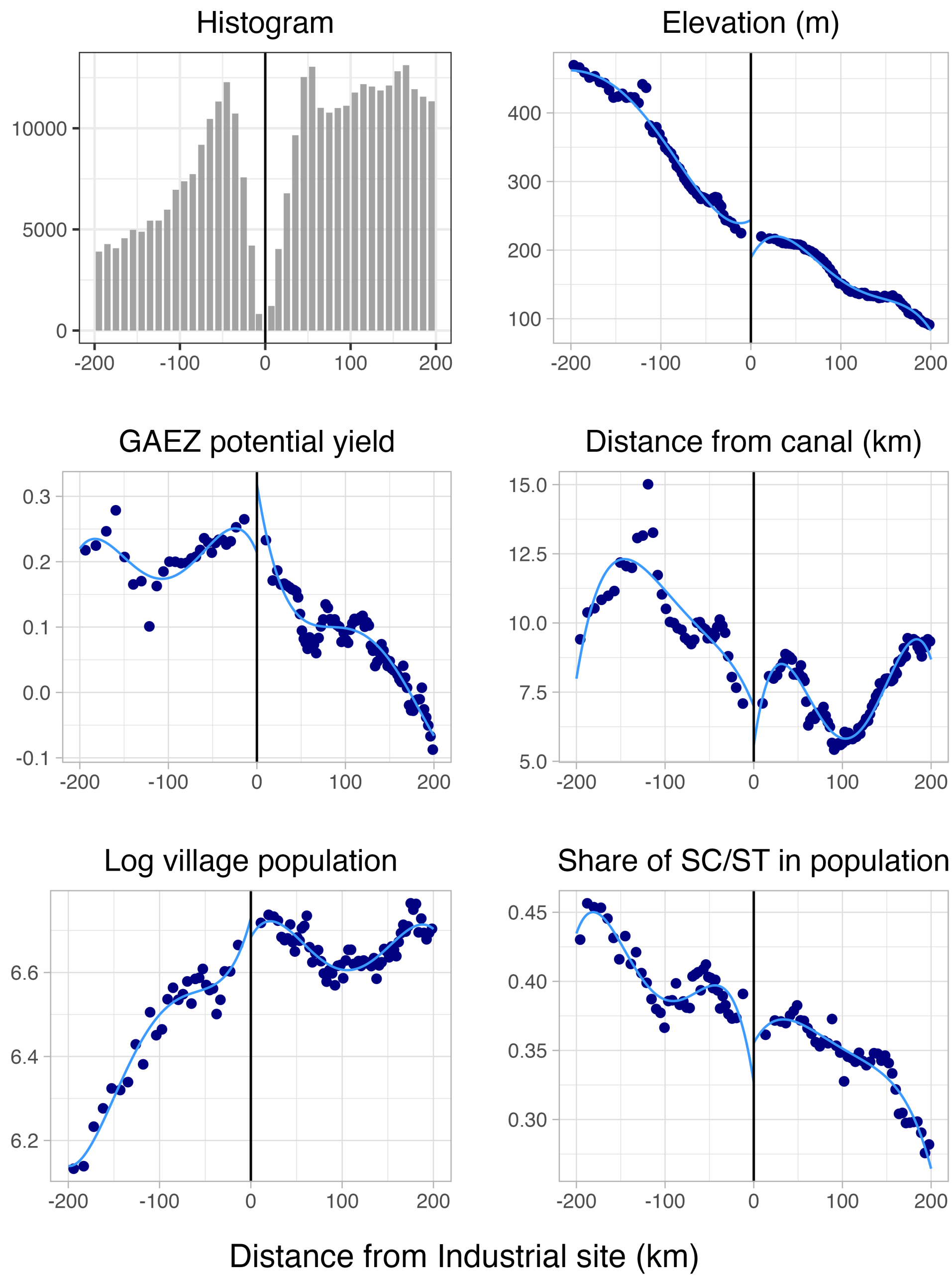

Figure 4: Continuity tests of a selection of covariates. The x-axis is geographical distance from a heavily-polluting industrial site. Areas with positive distance are downstream of the site; negative distance is upstream. Dots are binned scatterplots, showing means of each variable within quantiles of distance, adjusted for site fixed effects. Global polynomials are fitted separately on each side of the graph. GAEZ potential yield is normalized mean yield for all crops.

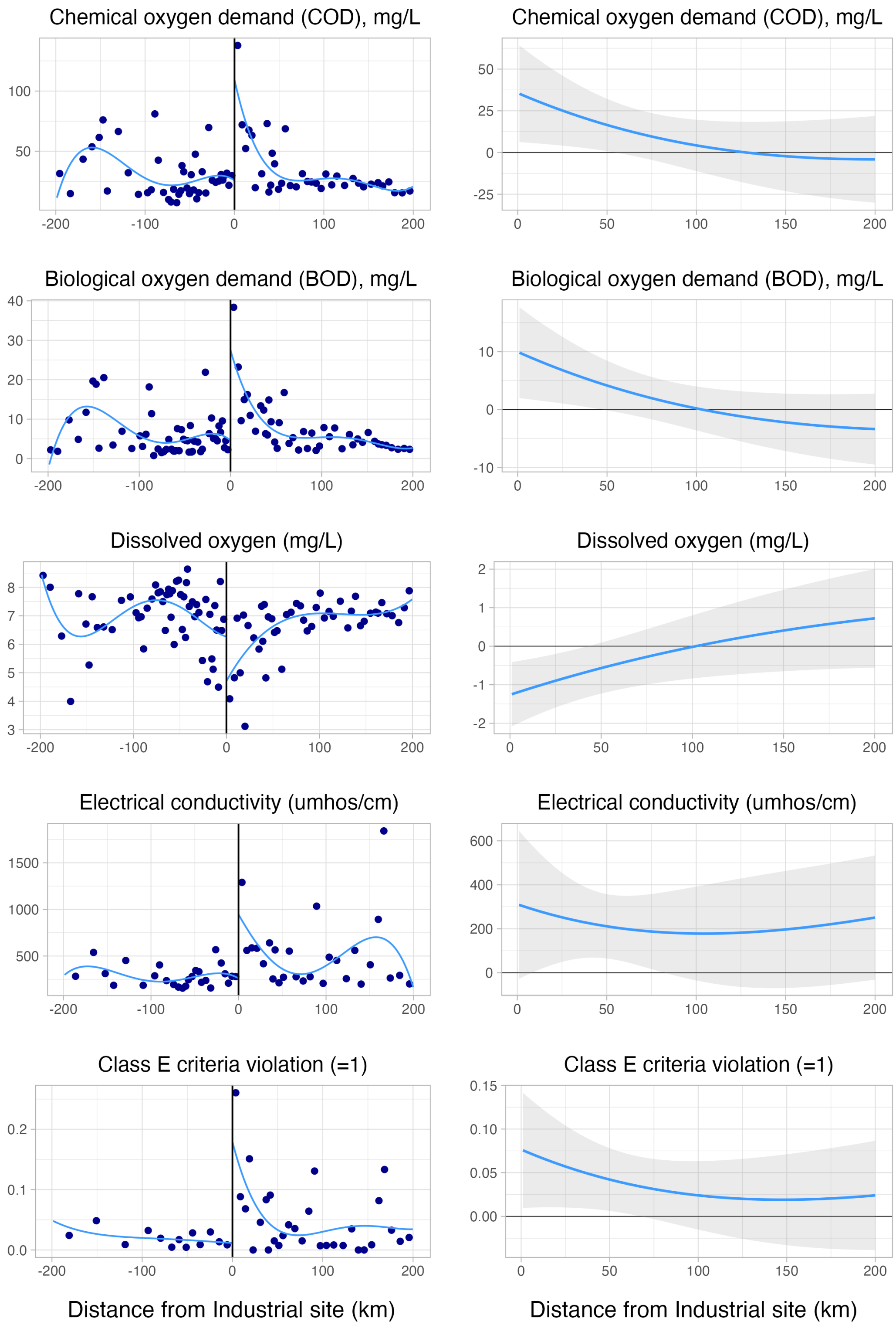

Figure 5: RD plots for surface water pollution measurements. Graphs on the left plot quantile-binned means and global polynomial fits. Positive distance indicates a monitoring station is downstream of the site; negative is upstream. Graphs on the right plot estimated impulse response functions (with 95% confidence intervals), showing how pollution concentrations decay downstream of an industrial site.

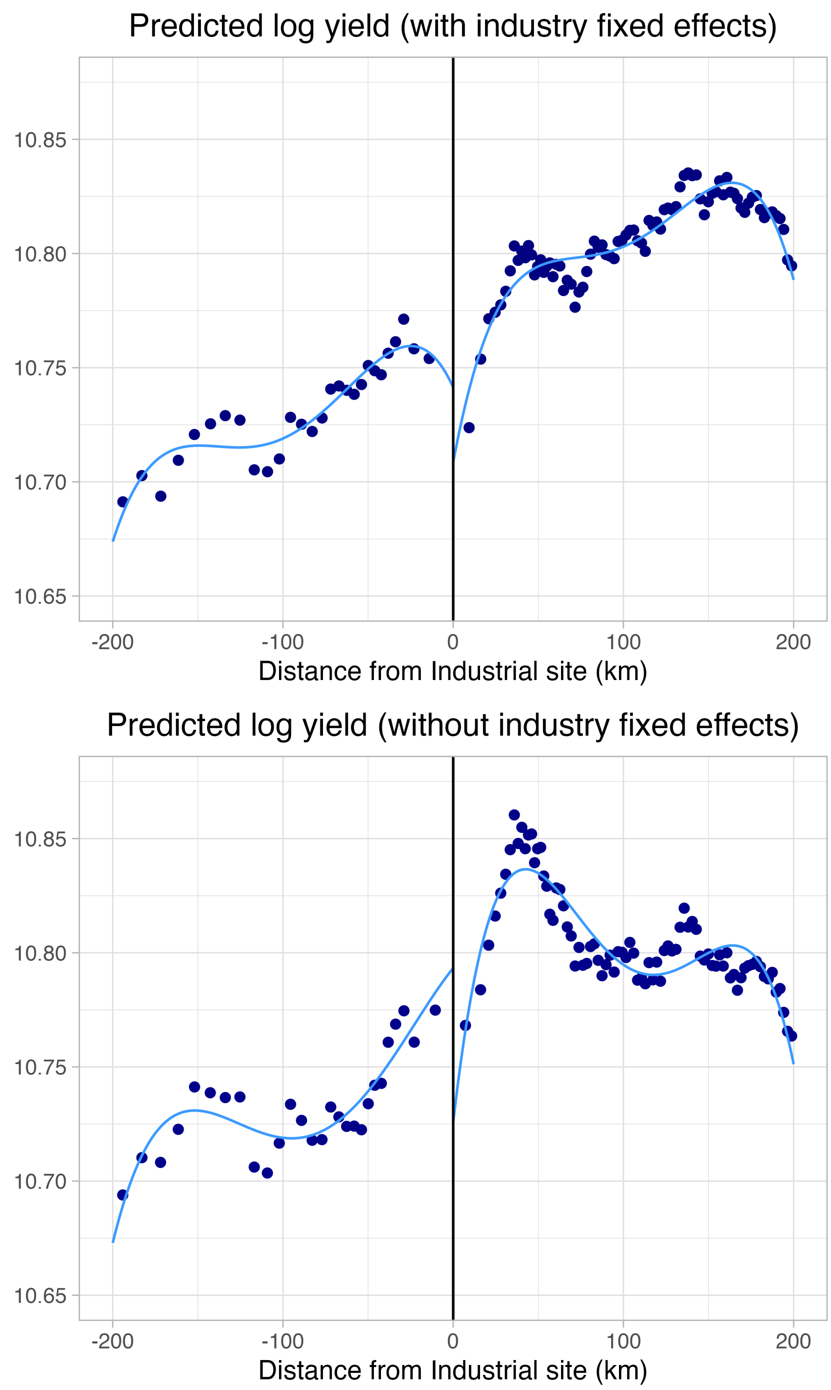

Figure 6: RD plots for crop yields as predicted from satellite data. The x-axis is geographical distance from a heavily-polluting industrial site. Positive distance indicates a village is downstream of the site; negative distance is upstream. Dots are binned scatterplots, showing means within quantiles of the running variable. Global polynomials are fitted separately on each side of the graph.

Tables

Table 1: Predictive Models of Crop Yields Using Satellite Data

Model	RMSE	R2
A. Candidate models, performance in test set		
1. Elastic net	0.523	0.123
2. Random forest	0.484	0.259
3. Boosted trees	0.499	0.205
4. Random forest, using raw bands only	0.511	0.170
5. Random forest, using VIs only	0.513	0.160
B. Alternative proxies, performance in test set		
6. Regression on log(Max NDVI - Min NDVI)	0.538	0.073
7. Regression on district-level VIs	0.540	0.048
C. Chosen model, performance in evaluation set		
3. Random forest	0.465	0.250

Notes: Performance of predictive models of observed crop yields using satellite data. Models are trained and evaluated on village-level data, calculated by averaging across sampled plots. Predictors are the village means of the annual maximum and minimum values of satellite bands and vegetation indices from Sentinel-2 after applying cloud and cropland masks. The outcome variable is the log of crop yields per hectare from sampled plots in each village, summed across crops (weighting by time-invariant average prices), and averaged across plots (weighting by plot area). The exception is Model 7, which is trained on district-level aggregate crop yield data and uses district-level vegetation indices as predictors but is evaluated on the same village-level data.

Table 2: Summary Statistics

Variable	Mean	SD	Observations
Surface water quality parameters			
Chemical Oxygen Demand (mg/L)	29.59	47.95	8790
Biological Oxygen Demand (mg/L)	5.786	12.55	10860
Dissolved Oxygen (mg/L)	7.116	1.617	10719
Electrical conductivity (μmhos/cm)	489	1647	9997
Class E criteria violated?	0.045	0.207	10890
Ground water quality parameters			
Electrical conductivity (μmhos/cm)	1178	911.8	37466
Total dissolved solids (mg/L)	611	566.4	8343
High pollution indicator	0.424	0.494	37960
Class E criteria violated?	0.102	0.302	35451
Irrigation sources			
Has a canal?	0.082	0.275	655694
Within 5 km of river?	0.102	0.303	655694
Shallow water table?	0.195	0.396	655694
Crop yield and agricultural inputs			
Predicted log revenue value of yield	10.77	0.246	605386
Crop area as share of village area	0.61	0.307	644254
Irrigated area as share of crop area	0.57	0.391	580560
Irrigation share from canals	0.176	0.321	584794
Irrigation share from wells	0.305	0.367	583884
Irrigation share from tanks or lakes	0.029	0.111	583872
Irrigation share from other sources (rivers)	0.053	0.182	585039
Socioeconomic outcomes			
Share of employment in agriculture	0.726	0.231	644580
Per capita consumption (1000s Rupees)	17.06	5.882	634663
Poverty rate	0.345	0.198	634663

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 various government agencies. The high pollution indicator and Class E criteria are described in section 5.1. The three irrigation source variables are described in section 6.2. Predicted yield is calculated from satellite data using machine learning as described in section 4. Agricultural inputs and employment share are from the Population Census of 2011. Per-capita consumption and poverty rate are from the Socio-economic and Caste Census of 2012.

Table 3: RD Estimates for Surface Water Quality

Dependent variable	Estimate	Robust 95% CI	p-value	Bandwidth	Effective N
Chemical Oxygen Demand (mg/L)	68.33	[56.96, 79.7]	0.000	62.8	2634
Biological Oxygen Demand (mg/L)	27.3	[26.61, 28]	0.000	35.2	1516
Dissolved Oxygen (mg/L)	-1.354	[-1.488, -1.22]	0.000	66.9	3290
Electrical conductivity (μmhos/cm)	1834	[1833, 1835]	0.000	14.1	650
Class E criteria violated?	0.243	[0.239, 0.246]	0.000	17.2	791

Notes: Geographic regression discontinuity estimates of the effects of severely-polluting industrial sites on water pollution in nearby rivers, immediately downstream of the sites. Estimates use local linear regression in geographical distance with site-by-year fixed effects, a triangular kernel, and an estimate-specific MSE-optimal bandwidth chosen using the algorithm of Calonico et al (2020). We report the bias-robust 95% confidence intervals and corresponding p-values of Calonico et al (2020), clustering by monitoring station. Effective N is the number of observations that fall within the bandwidth and are therefore used in estimation.

Table 4: RD Estimates for Ground Water Quality

Dependent Variable	Estimate	Robust 95% CI	p-value	Bandwidth	Effective N
Panel A: Main RD effect					
Electrical conductivity	108	[15.88, 201]	0.022	51.9	6610
Total dissolved solids	-1.28	[-115.4, 113]	0.983	90.3	2292
High pollution indicator	0	[-0.041, 0.041]	0.993	56.8	7681
Class E criteria violated?	0.016	[-0.012, 0.045]	0.258	51.0	6110
Panel B: Has a canal?					
Electrical conductivity	-701	[-1148, -253]	0.002	20.7	253
Total dissolved solids	399	[189.7, 608]	0.000	45.4	133
High pollution indicator	0.515	[0.205, 0.824]	0.001	19.3	221
Class E criteria violated?	0.029	[-0.07, 0.127]	0.571	16.6	146
Panel C: Within 5 km of river?					
Electrical conductivity	700	[571.8, 828]	0.000	49.8	1576
Total dissolved solids	210	[74.98, 345]	0.002	98.5	654
High pollution indicator	0.07	[0.015, 0.125]	0.013	76.3	2755
Class E criteria violated?	0.07	[0.04, 0.101]	0.000	88.6	2899
Panel D: Shallow water table (<8m)?					
Electrical conductivity	696	[443.7, 948]	0.000	18.3	390
Total dissolved solids	290	[39.29, 541]	0.023	41.3	277
High pollution indicator	0.598	[0.503, 0.693]	0.000	16.6	331
Class E criteria violated?	0.028	[-0.014, 0.07]	0.194	28.1	795

Notes: Geographic regression discontinuity estimates of the effects of severely-polluting industrial sites on groundwater pollution in villages immediately downstream of the sites; see notes to Table 3. Panels B-D restrict the sample to villages most likely affected by specific physical pathways of pollution transport, as described in section 6.2. Sample includes villages within 20 km of a flow path that passes near each industrial site, as defined in section 3.1. Inference is clustered by subdistrict. ↩

Table 5: RD Estimates for Predicted Crop Yield

Dependent variable: *Predicted Log Revenue Value of Yield*					
	Estimate	Robust 95% CI	p-value	Bandwidth	Effective N
Panel A: Main Effect					
Downstream effect	-0.027	[-0.071, 0.016]	0.220	55.1	80694
Panel B: Robustness to controling for distance to river					
Downstream effect	-0.031	[-0.074, 0.012]	0.161	48.7	64785
Panel C: Robustness to controling for pre-treatment variables					
Downstream effect	-0.027	[-0.07, 0.016]	0.215	54.9	80072
Panel D: Robustness to controling for irrigation dummies					
Downstream effect	-0.025	[-0.068, 0.018]	0.250	55.1	80576

Notes: Geographic regression discontinuity estimates of the effects of severely-polluting industrial sites on crop yield in villages immediately downstream of the sites; see notes to Table 3. The outcome variable, log crop yield per hectare, is predicted from satellite data using machine learning as described in section 4. Sample includes villages within 20 km of a flow path that passes near each industrial site, as defined in section 3.1. Inference is clustered by subdistrict.

Table 6: RD Estimates for Agricultural Inputs

Dependent variable	Estimate	Robust 95% CI	p-value	Bandwidth	Effective N
Crop area as					
Share of village area	-0.026	[-0.08, 0.028]	0.352	54.3	37605
Irrigated area as					
Share of crop area	0.057	[-0.004, 0.119]	0.067	41.7	25763
Irrigation share from					
Canals	0.005	[-0.034, 0.044]	0.809	47.8	30516
Wells	0.022	[-0.026, 0.069]	0.372	41.2	25610
Tanks or lakes	-0.01	[-0.032, 0.013]	0.401	50.7	32308
Other (rivers)	-0.003	[-0.018, 0.011]	0.649	89.6	57851
Socio-economic outcomes					
Share of employment in ag	-0.008	[-0.06, 0.045]	0.767	69.5	48349
Per capita consumption (Rupees)	-484	[-1212, 244]	0.192	102.0	72305
Poverty rate	0.003	[-0.019, 0.026]	0.773	79.4	56365

Notes: Geographic regression discontinuity estimates of the effects of severely-polluting industrial sites on agricultural inputs in villages immediately downstream of the sites; see notes to Table 5. Outcome variables are from the Population Census of 2011, except per-capita consumption and poverty rate are from the Socio-Economic and Caste Census of 2012.

Table 7: RD Estimates for Yield and Inputs: Heterogeneity by Irrigation Sources

Sample restriction	Estimate	Robust 95% CI	p-value	Bandwidth	Effective N
Outcome: Predicted log revenue value of yield					
Has a canal?	-0.105	[-0.199, -0.012]	0.027	53.1	3035
Within 5 km of river?	-0.035	[-0.088, 0.017]	0.190	55.6	10193
Shallow water table?	-0.045	[-0.16, 0.069]	0.440	45.2	7756
Outcome: Crop area as share of village area					
Has a canal?	0.031	[-0.079, 0.141]	0.578	43.3	2310
Within 5 km of river?	0.028	[-0.039, 0.095]	0.407	60.0	10558
Shallow water table?	0.057	[-0.058, 0.171]	0.331	52.9	8557
Outcome: Irrigated area as share of crop area					
Has a canal?	0.035	[-0.115, 0.185]	0.645	35.0	1681
Within 5 km of river?	0.048	[-0.011, 0.106]	0.111	46.3	7288
Shallow water table?	0.039	[-0.056, 0.134]	0.417	46.4	6682
Outcome: Share of irrigation from canals					
Has a canal?	0.013	[-0.105, 0.131]	0.827	34.7	1667
Within 5 km of river?	-0.01	[-0.059, 0.04]	0.695	68.4	11240
Shallow water table?	0.005	[-0.052, 0.062]	0.869	32.7	4554
Outcome: Share of irrigation from wells					
Has a canal?	0.002	[-0.059, 0.063]	0.946	34.9	1675
Within 5 km of river?	0.013	[-0.038, 0.064]	0.627	49.3	7934
Shallow water table?	0.015	[-0.025, 0.055]	0.470	41.4	6006
Outcome: Per capita consumption (Rupees)					
Has a canal?	117	[-1920, 2155]	0.910	43.2	2270
Within 5 km of river?	315	[-784.9, 1415]	0.575	77.2	14171
Shallow water table?	-778	[-2811, 1255]	0.453	58.4	10037

Notes: Geographic regression discontinuity estimates of the effects of severely-polluting industrial sites on predicted crop yield and agricultural inputs in villages immediately downstream of the sites, for sub-samples restricted to villages most likely affected by specific physical pathways of pollution transport. See notes to Table 5.

References

Abdoli, Majid. 2022. Sustainable Use of Sewage Sludge in Soil Fertility and Crop Production. In Sustainable Management and Utilization of Sewage Sludge, eds. Vishnu D. Rajput, Ajar Nath Yadav, Hanuman Singh Jatav, Satish Kumar Singh, and Tatiana Minkina. Cham: Springer International Publishing, 297–333.

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

Ahmed, Jebin, Abhijeet Thakur, and Arun Goyal. 2021. Industrial Wastewater and Its Toxic Effects. In Biological Treatment of Industrial Wastewater. The Royal Society of Chemistry.

Alix-García, Jennifer, and Daniel L. Millimet. 2023. Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation. Journal of the Association of Environmental and Resource Economists 10 (5): 1335–1367.

Amale, Hardeep Singh, Pratap Singh Birthal, and Digvijay Singh Negi. 2023. Delayed monsoon, irrigation and crop yields. Agricultural Economics n/a (n/a).

Andarge, Tihitina. 2020. Effect of Incomplete Information on Ambient Pollution Levels.

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): 257–280.

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.

Bajpai, Pratima. 2013. Pulp Bleaching and Bleaching Effluents. In Bleach Plant Effluents from the Pulp and Paper Industry, ed. Pratima Bajpai. Heidelberg: Springer International Publishing, 13–19.

Baylis, Kathy, Thomas Heckelei, and Hugo Storm. 2021. Machine learning in agricultural economics. In Handbook of Agricultural Economics, volume 5. Elsevier, 4551–4612.

Bedane, Dejene Tsegaye, and Seyoum Leta Asfaw. 2023. Microalgae and co-culture for polishing pollutants of anaerobically treated agro-processing industry wastewater: The case of slaughterhouse. Bioresources and Bioprocessing 10 (1): 81.

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): 2189–2194.

Burney, Jennifer, and V. Ramanathan. 2014. Recent climate and air pollution impacts on Indian agriculture. Proceedings of the National Academy of Sciences 111 (46): 16319–16324.

Calonico, Sebastian, Matias D Cattaneo, and Max H Farrell. 2020. Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. The Econometrics Journal 23 (2): 192–210.

Cattaneo, Matias D., Nicolas Idrobo, and Rocio Titiunik. 2024. A Practical Introduction to Regression Discontinuity Designs: Extensions.

Central Pollution Control Board. 2009. Criteria for comprehensive environmental assessment of industrial clusters. Technical report.

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): 12936–12941.

Colmer, Jonathan. 2021. Temperature, Labor Reallocation, and Industrial Production: Evidence from India. American Economic Journal: Applied Economics 13 (4): 101–124.

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): 306–325.

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.

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 Organization.

Flynn, Patrick, and Michelle M. Marcus. 2021. A Watershed Moment: The Clean Water Act and Infant Health.

Garcia, Alberto, and Robert Heilmayr. 2022. Conservation Impact Evaluation Using Remotely Sensed Data.

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: 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): 709–715.

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.

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

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

Haseeb, Muhammad. 2024. Resource Scarcity and Cooperation.

Hawkins, Gary L., and L. Mark Risse. 2017. Beneficial Reuse of Municipal Biosolids in Agriculture.

He, Guojun, Shaoda Wang, and Bing Zhang. 2020. Watering Down Environmental Regulation in China. The Quarterly Journal of Economics 135 (4): 2135–2185.

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.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. Misallocation and Manufacturing TFP in China and India. The Quarterly Journal of Economics 124 (4): 1403–1448.

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 .

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): 349–396.

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.

Kosmowski, Frederic, Jordan Chamberlin, Hailemariam Ayalew, Tesfaye Sida, Kibrom Abay, and Peter Craufurd. 2021. How accurate are yield estimates from crop cuts? Evidence from smallholder maize farms in Ethiopia. Food Policy 102: 102122.

Lindhjem, Henrik, Tao Hu, Zhong Ma, John Magne Skjelvik, Guojun Song, Haakon Vennemo, Jian Wu, et al. 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): 202–219.

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 .

Murty, M.N., and Surender Kumar. 2011. Water pollution in india: An economic appraisal. In India infrastructure report.

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

National Academies. 1996. Use of Reclaimed Water and Sludge in Food Crop Production. Washington, D.C.: National Academies Press.

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): 12302–12308.

Proctor, Jonathan, Tamma Carleton, and Sandy Sum. 2023. Parameter Recovery Using Remotely Sensed Variables.

Rai, Prabhat Kumar, Sang Soo Lee, Ming Zhang, Yiu Fai Tsang, and Ki-Hyun Kim. 2019. Heavy metals in food crops: Health risks, fate, mechanisms, and management. Environment International 125: 365–385.

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

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): 547–560.

Scott, C.I., N. I. Faruqui, and L. Raschid-Sally. 2004. Wastewater use in irrigated agriculture: Confronting the livelihood and environmental realities.

Sekhri, Sheetal. 2014. Wells, Water, and Welfare: The Impact of Access to Groundwater on Rural Poverty and Conflict. American Economic Journal: Applied Economics 6 (3): 76–102.

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): 317–322.

Sudarshan, Shanmugam, Sekar Harikrishnan, Govindarajan RathiBhuvaneswari, Venkatesan Alamelu, Samraj Aanand, Aruliah Rajasekar, and Muthusamy Govarthanan. 2023. Impact of textile dyes on human health and bioremediation of textile industry effluent using microorganisms: Current status and future prospects. Journal of Applied Microbiology 134 (2): lxac064.

Taylor, Charles A., and Hannah Druckenmiller. 2022. Wetlands, Flooding, and the Clean Water Act. American Economic Review 112 (4): 1334–1363.

Vyas, Ananya. 2022. Explainer: What is causing the mass death of fish in India's water bodies?

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): 471–473.

Weiss, Marie, Frédéric Jacob, and Grgory Duveiller. 2020. Remote sensing for agricultural applications: A meta-review. Remote sensing of environment 236: 111402.

World Bank, and State Environmental Protection Administration. 2007. Cost of pollution in China: Economic estimates of physical damages. Technical Report 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.

Zhang, Xiaowei, and Qian Lu. 2024. Cultivation of microalgae in food processing effluent for pollution attenuation and astaxanthin production: a review of technological innovation and downstream application. Frontiers in Bioengineering and Biotechnology 12.