Online Appendix to “Industrial Water Pollution and Agricultural Production in India”

Nick Hagerty
Anshuman Tiwari *

15 December 2024

1 Appendix

1.1 Details of hydrological modeling

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

Flow length raster. We obtain a digital elevation model (DEM) at 15 arc-second resolution for the South Asia area from the HydroSHEDS project of the United States Geological Survey. From this DEM, we use the Spatial Analyst tools in ArcGIS Pro to fill sinks, create a flow direction raster (using the D8 method), and derive a flow length raster. This raster gives the distance along rivers that a particle released at each cell must travel to reach the ocean (or the edge of the raster).

Sample construction. To define the sample of villages for each industrial site, we first create a reference flow line. We use the Trace Downstream tool in ArcGIS Pro to find the site’s flow path, i.e., the route that effluent released at the site must follow to reach the ocean. We then find the point on this flow path that is 25 km downstream of the site (the lower yellow dot in Figure 2 of the main paper). Next, we use the Watershed tool in ArcGIS Pro to find the area that drains into that point. We find the flow lengths of all villages within this watershed by intersecting the watershed polygon with village centroids and matching village centroids to the flow length raster. We identify the longest possible flow path within this watershed by choosing the village at the 95th percentile of flow length within this set (the upper yellow dot). We use the 95th percentile instead of the maximum to avoid erroneous values that sometimes arise at the edges of watershed polygons. Finally, to define the sample, we find the flow path of the chosen furthest-upstream village, generate a 20-km buffer around each flow path, and intersect this buffer with centroids of villages and monitoring stations.

*Hagerty: Montana State University (email: nicholas.hagerty@montana.edu); Tiwari: Energy Policy Institute at the University of Chicago (email: atiwari2@uchicago.edu).

1.2 Impulse response functions

To estimate the non-local effects of industrial sites (i.e., further downstream than the RD cutoff), we use models of the following form:

$$y_{ist} = \gamma Distance_{is} + f(Distance_{is} \times Downstream_{is}) + \alpha_{st} + \varepsilon_{ist} \tag{1}$$

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

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

The assumption required for the spatial response function is considerably stronger than for the RD design. This design requires that the upstream trend can be extrapolated – that without the industrial sites, outcomes would have continued to follow the upstream trend downstream for as far as we estimate the function. This assumption is most likely to hold nearest to the downstream cutoff, so the function is less reliable the further downstream we go. Despite these limitations, this design is the best available method to estimate the effects of industrial clusters away from the cutoff.

1.3 Details of predictive modeling of crop yields

1.3.1 Survey versus satellite data

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

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

1.3.2 Vegetation indices

The remote sensing literature has proposed a number of measures to proxy for crop yields, called vegetation indices (VIs). Rather than choose from among them, we use all VIs that can be calculated from available data, following Lobell et al. (2020). We use these VIs as predictors in our model, as well as the raw variables (bands) used to calculate them. We use both types of predictors because the theoretically-grounded VIs may provide helpful structure to fit the model, while their underlying bands allow more flexibility.

All VIs aim to capture the amount of photosynthetic activity in plants, which correlates with yields. Chlorophyll, the pigment that gives leaves their green color, absorbs much of the red light in the visible spectrum in healthy plants. Other cell structures of the plant reflect most of the near-infrared light in the invisible part of the electromagnetic spectrum. A healthy plant with high photosynthetic activity due to high amounts of chlorophyll will reflect less red light and more near-infrared light. Like cameras, satellite instruments capture the amount of light reflected in these different bands of the electromagnetic spectrum. Each VI is a function of the values recorded in different bands.

NDVI and EVI are the two indices most commonly used in the scientific literature to proxy for agricultural output. NDVI uses red and near-infrared light; EVI is similar but uses additional information from the blue part of the electromagnetic spectrum to reduce atmospheric interference and the influence of background vegetation (Son et al. 2014). The other four VIs are variations on the same idea; each has been shown to be useful in different settings (Burke and Lobell 2017).

1.3.3 Satellite data

We extract minimum and maximum values of each VI during agricultural years 2015-17 from the Sentinel-2 MSI satellite[1] and aggregate them to villages.[2] Maximum values of VIs are often found to be most strongly predictive of crop yields; minimum values (which likely occur during the off-season) may help control for background factors related to land cover (Asher and Novosad 2020). India's agricultural year spans July 1 of the reference year through June 30 of the following year. We use years 2015-17 for model training to correspond to the availability of both Red-Edge bands (to calculate NDVI705 and NDVI740) and village-level training data on crop yield.

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

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

[2] Sentinel-2 is a satellite launched by the European Space Agency that records images at each point on Earth's land surface approximately once every 10 days, at a spatial resolution of 10 to 60 meters depending on band. The other major source of publicly available satellite imagery, NASA's Landsat 7, does not measure wavelengths in the ranges required to calculate NDVI705 and NDVI740.

1.3.4 Village-level crop yields.

The Cost of Cultivation survey is implemented by the Directorate of Economics and Statistics, Ministry of Agriculture.[3] The survey design follows a three-stage stratified random sampling with the subdistrict (aka tehsil, the third administrative tier below state and district) as the first stage unit, a cluster of villages as the second stage unit and an operational holding as the third and ultimate stage unit. Data is collected by major agricultural universities by asking each sample household to complete detailed daily records. The survey is conducted across 16 agriculturally important states, which together account for 99% of cropped area in India, as calculated from the Population Census of 2011.

This survey is by far the largest nationally-representative source of publicly available microdata for India that we are aware of. The data contain plot-level information on crop type, area planted, output, and earnings for 29 major crops in approximately 3000 villages in a rotating 3-year panel between 2008 and 2019. We merge villages to other datasets by geolocating them using ArcGIS, confirming accuracy for a random sample, and spatially joining them to village boundaries. A spatial join is necessary because a merge on village name performs poorly, due to inconsistency in the spelling of village names and changes in village definitions over time.

For each village, we calculate crop yields per hectare, averaged across sampled plots, for agricultural years 2015-17. To compare across crops, we weight by prices, so we also refer to the outcome variable as the revenue value of yield. For price weights, we calculate the average revenue per unit of physical output for each crop in each district using all years in the data. These weights are time-invariant, so they can be thought of as the expected price for each crop, and they allow us to estimate impacts to production excluding equilibrium price changes. We then take the mean of price-weighted yields across all sampled plots in each village, weighting by share of crop area in each plot. Our final outcome variable is the log of this value.

1.3.5 Model training

We use the `glmnet`, `ranger`, and `lightgbm` packages in R. We tune the hyperparameters of each model through 3-fold cross-validation with 3 repeats. Where possible (i.e., for all models but boosted trees), we weight observations by the number of sampled plots in each village in both tuning and fitting.

We conduct initial tuning to identify hyperparameters whose optimal values appear independent of others' values, fix those ones, and further tune the others. For the elastic net, we tune the penalty and fix the mixture at 0. For the random forest, we tune the number of predictors to be sampled at each split and the minimum number of data points for a node to be split further and fix the number of trees at 1000. For boosted trees, we tune the same hyperparameters as the random forest plus the maximum depth of the tree and fix the learning rate at 0.1 and the number of trees at 50.

In general, VIs are just functions of the bands, so in principle a sufficiently flexible predictive model should be able to learn this relationship. However, we calculate the VIs before taking annual minimum and maximum values and then aggregating to villages, so in our data the VI and band predictors are not related by simple transformations.

[3] https://mospi.gov.in/412-cost-cultivation-principal-crops

1.4 District-level predictive model

In addition to our village-level predictive model, we build and evaluate a predictive model trained on district-level crop yield data.

Satellite data. To calculate district-level VIs, we take means of village-level values, weighting villages by agricultural land area from the population census.

District-level crop yields. We calculate price-weighted crop yields from the District Level Database compiled by ICRISAT.[4] This data contains information on crop area planted, output, and prices for 16 major crops, for 571 districts across 20 states from 1990-2017. Price data covers about 79% of all area under cultivation. Revenue value of yield is calculated by multiplying the quantity of each crop by the (time-invariant) mean price for that crop in that district between 1990-2002. For districts without price data, we impute the state mean if available or the national mean otherwise.

Results. We first verify that our calculated VIs are individually predictive of, and positively correlated with, crop yields. To do so, we regress log revenue value of yield on the log of the difference between maximum and minimum values of each VI, following Asher and Novosad (2020). This is a district-level cross-sectional regression; we omit spatial fixed effects since our final research design relies on spatial variation. Results are shown in the first four columns of Table 2. NDVI, EVI, GCVI, and MTCI are each positively correlated with log yields and individually explain a substantial fraction (between 6 and 21 percent) of the in-sample variation in district-level log yields.

Next, we fit our predictive model: We regress log revenue value of yield on all six VIs, with maximum and minimum values entering separately and linearly, following Lobell et al. (2020). Results are shown in column (5) of Table 2. Individual coefficients lack an intuitive interpretation, since each is conditional on all the others. The explanatory power of this regression far exceeds any of the individual VIs, with an R^2 of 0.39. However, this result represents in-sample performance. We evaluate the model's out-of-sample performance in village-level data in Section 4 of the main paper.

1.5 Details of village covariates and boundaries

For baseline village covariates, we use the Population Census of 2001, which includes many variables on population, employment, amenities, and infrastructure. For agricultural inputs and village outcomes, we use the Population Census of 2011, since it is collected closer to the time period of our crop yield data. We obtain cleaned Population Census data along with geospatial data on village boundaries from NASA's Socioeconomic Data and Applications Center.[5]

Because villages and towns sometimes split or merge, we use consistent definitions from the Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG) provided by the Development Data Lab.[6] The SHRUG provides an identifier called a shrid for a group of

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

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

[6] https://www.devdatalab.org/shrug_download/

contiguous villages or towns that can be combined into unchanged spatial entities over several decades.[7] For employment in polluting industries, we construct the village sum of firm-level employment from the Economic Census of 2013 within sectors that are classified by the CPCB as major water polluting industries.[8]

For cropland quality and crop suitability, we use potential yields from the Global Agro-Ecological Zones (GAEZ) database of the Food and Agricultural Organization (FAO).[9] This database provides crop-specific yields that are agronomically-possible upper limits under local agro-climatic, soil and terrain conditions and with specific farm management and agronomic input levels. We obtain this data at the village level from SHRUG.

1.6 Heterogeneous effects by industrial site characteristics

Here we attempt to test the hypothesis that the effects of industrial effluent on crop yields are small because the effluent contains beneficial nutrients in addition to harmful pollutants.

Treatment effects of specific pollutants are difficult to estimate reliably, because they are numerous, highly correlated with each other, and the available water quality data lacks spatial density. Instead, we estimate effects of different groupings of industrial sites on crop yields, using two approaches. The first approach is to use data on demographics in jurisdictions near the industrial site to proxy for the amount of "bad" industrial effluent versus "good" municipal effluent. The second approach is to use observed changes in nitrates downstream of the industrial site to proxy for the amount of beneficial nutrients released at the site from any source.

Appendix Table 3 reports results from two approaches. Based on point estimates alone, crop yield effects are concentrated among sites expected to have more industrial effluent relative to municipal effluent (Panel C),[10] and among sites that release relatively little nitrate (Panel D).[11] However, none of the estimates nor their differences are statistically significant.

[7] Almost 96% of villages from the 2001 population match a single shrid and do not require spatial aggregation. For the rest, we dissolve polygons boundaries to obtain shrid boundaries, and aggregate data over the villages within each shrid.

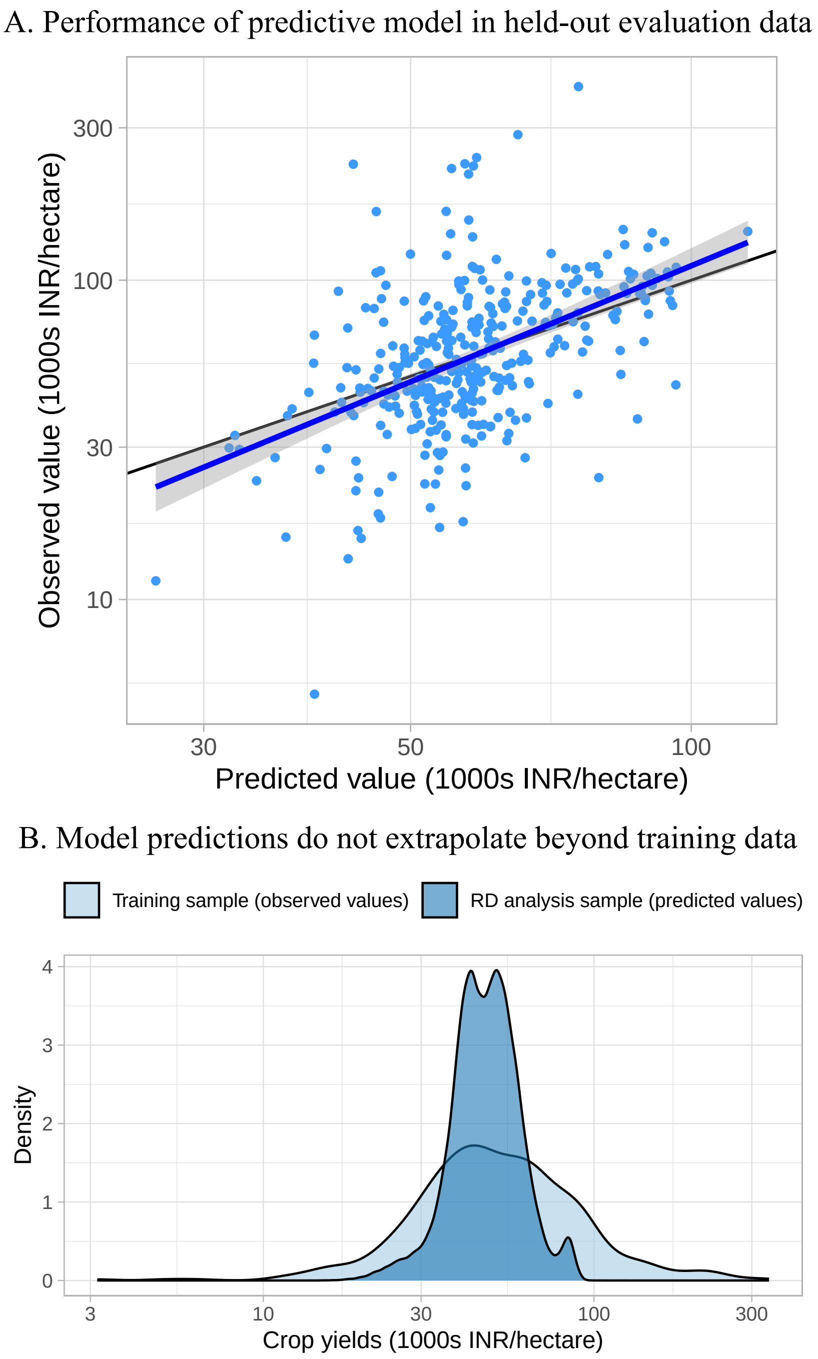
[8] The economic census covers the universe of non-farm establishments in India. The corresponding 3-digit industry codes in the census are manually coded to match major polluting sectors based on guidance available at https://cpcb.nic.in/faq.php (#31). We use the economic census for this purpose as it provides these 3-digit industry codes whereas the population census does not.

[9] https://gaez.fao.org/datasets/hqfao::gaez-agro-climatic-potential-yield/about.

[10] Based on the ratio of polluting-sector employment to population. We focus on this ratio because polluting-sector employment and population are highly correlated, so each likely confounds the other's effects. For reference, Panels A and B consider each variable separately.

[11] Since site-specific RDs are too noisy, we compute the downstream-upstream difference in mean nitrate observations within 100 km of the site.

2 Appendix Figures



A. Performance of predictive model in held-out evaluation data

B. Model predictions do not extrapolate beyond training data

Figure 1: Results of predictive model for village-level crop yields. Panel A plots observed values against predictions from the random forest model in the held-out evaluation set (a 20% random sample of the available Cost of Cultivation data); the model is fit on the rest of the data (the training + test sets). Panel B compares the density of crop yield values in the training sample (observed values) and the final sample for RD analysis (predicted values). The outcome 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).

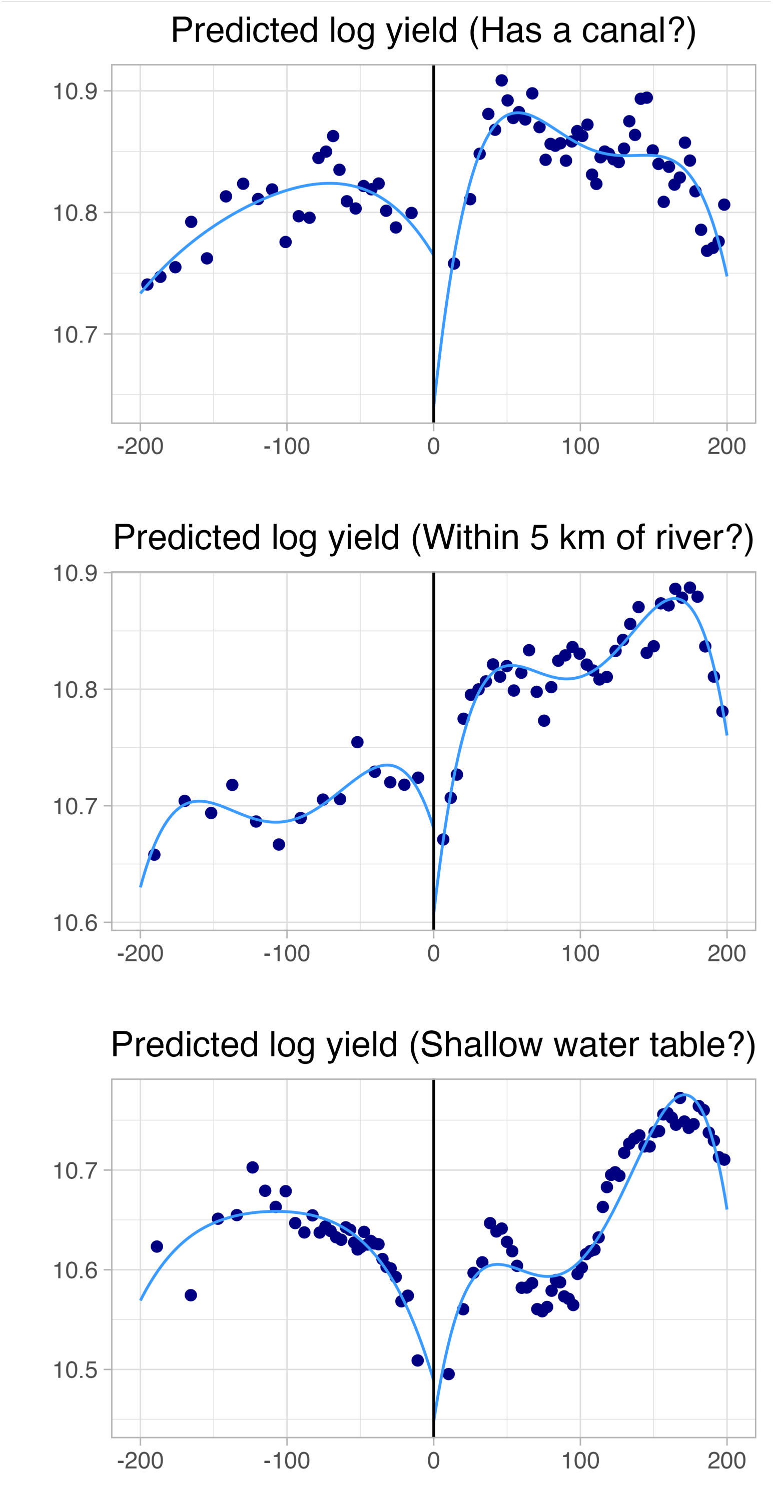


Figure 2: RD plots for crop yields as predicted from satellite data. Each graph restricts the sample to villages most likely to be affected by specific pathways of pollution transport.

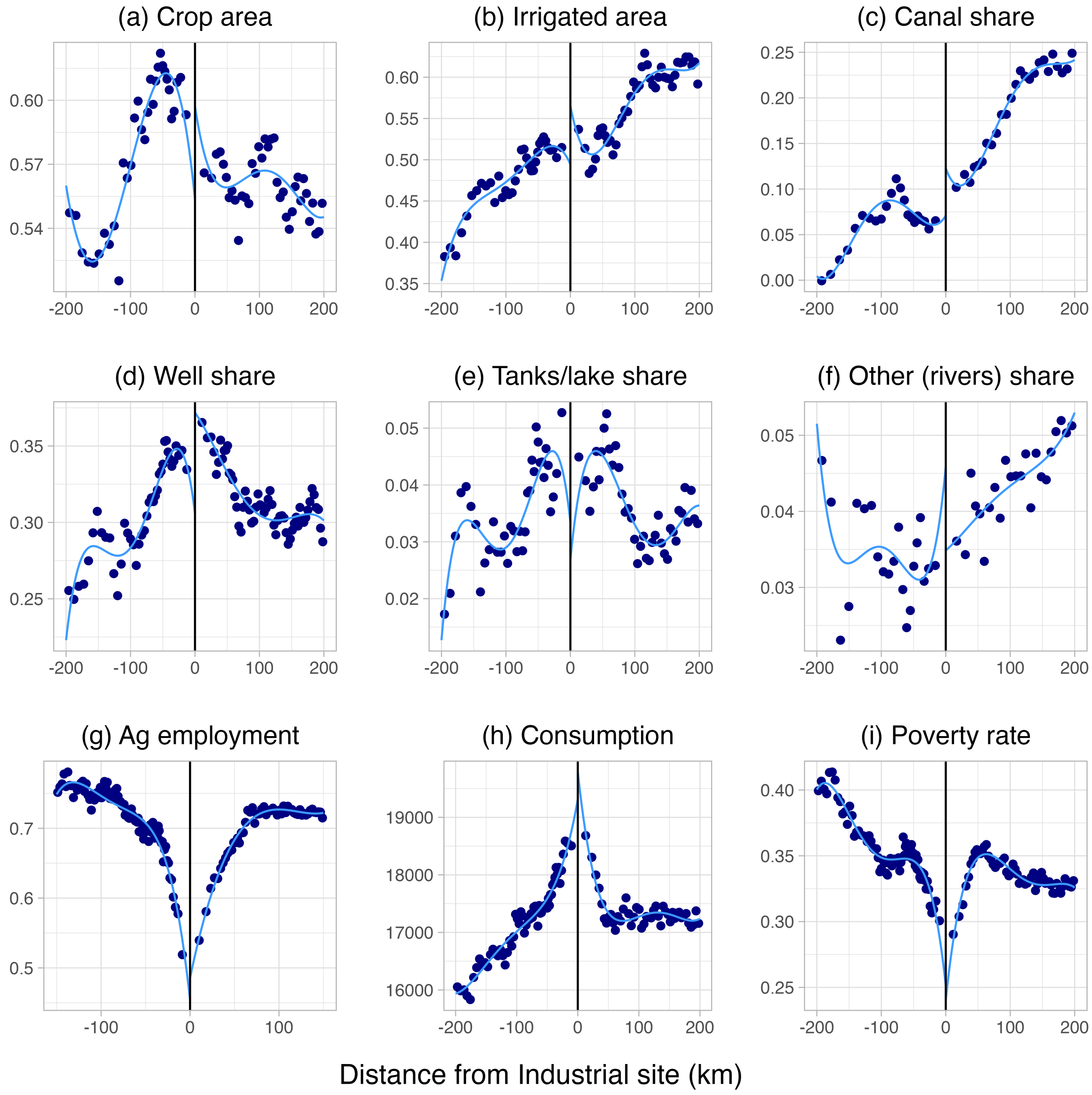


Figure 3: RD plots for agricultural inputs and economic outcomes. Outcome variables are: (a) crop area as a share of village area, (b) irrigated area as a share of crop area, (c)-(f) share of crop area irrigated from specific sources, (g) share of employment in agriculture, (h) per-capita consumption in rupees, and (i) poverty rate.

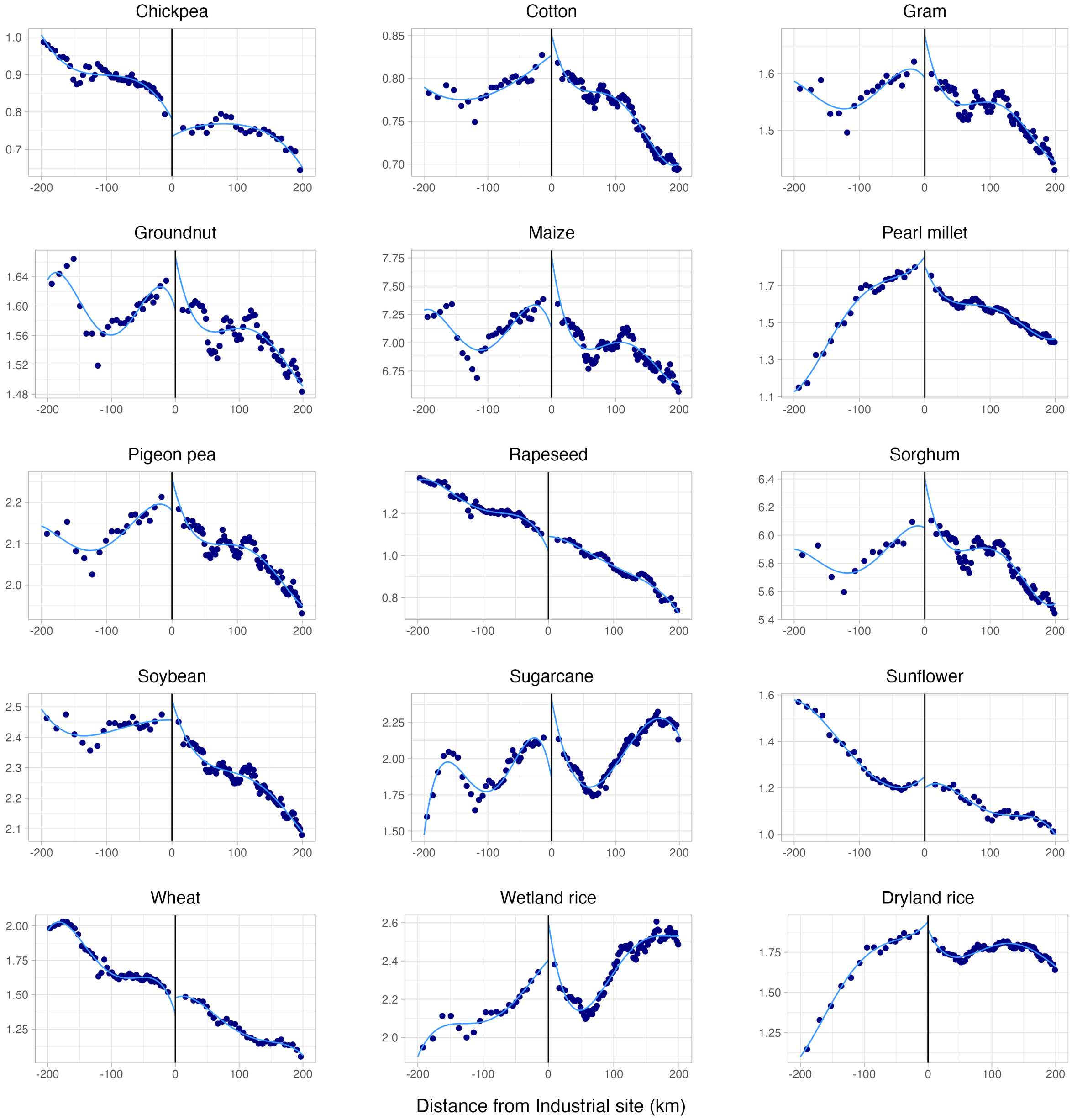


Figure 4: Continuity tests for potential yields of specific crops from GAEZ data.

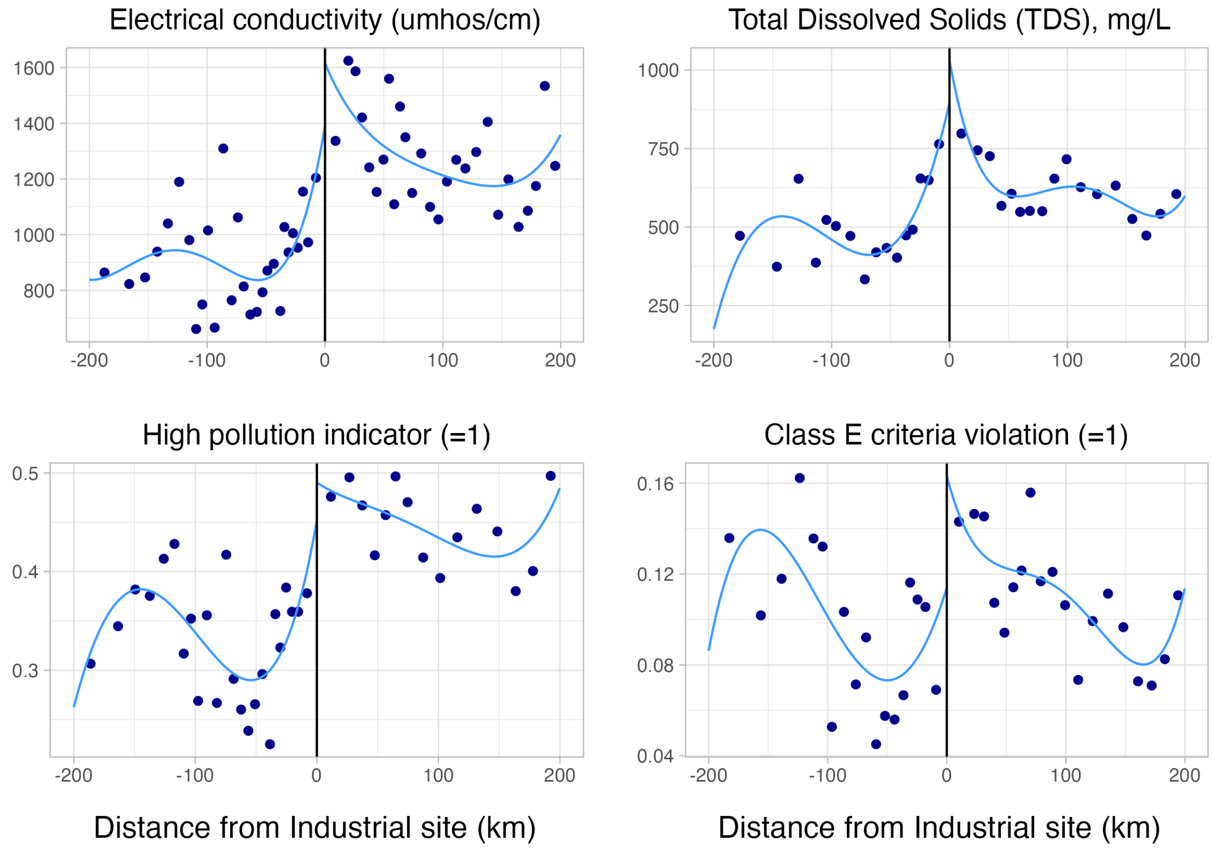


Figure 5: RD plots for groundwater pollution measurements. Graphs plot mean values of each parameter within quantile bins (and global polynomial fits) of distance from industrial site. Positive distance indicates a monitoring station is downstream of the site; negative is downstream.

3 Appendix Tables

Table 1: RD Estimates for Continuity of Covariates

Dependent Variable	Estimate	Robust 95% CI	p-value	Bandwidth	Effective N
Panel A: Physical Characteristics					
Distance from canal (km)	0.501	[-1.57, 2.57]	0.635	41.6	29357
Distance to nearest town (km)	0.856	[-2.1, 3.81]	0.570	45.7	32776
Elevation (m)	-3.49	[-13.3, 6.34]	0.487	31.8	20937
Panel B: GAEZ potential yield - High Input Scenario (kg/ha)					
Chickpea	0.035	[-0.009, 0.079]	0.123	38.6	26994
Cotton	0.007	[-0.018, 0.032]	0.580	48.0	34616
Dryland rice	0.035	[-0.034, 0.104]	0.317	52.8	38382
Gram	0.024	[-0.037, 0.085]	0.437	47.6	34326
Groundnut	0.005	[-0.058, 0.068]	0.884	43.0	30504
Maize	0.109	[-0.124, 0.341]	0.360	61.0	44582
Pearl millet	0.045	[-0.03, 0.121]	0.236	43.4	30880
Pigeon pea	0.028	[-0.054, 0.11]	0.499	49.0	35404
Rapeseed	0.025	[-0.017, 0.068]	0.237	36.1	24869
Sorghum	0.091	[-0.108, 0.29]	0.369	50.0	36176
Soybean	0.063	[-0.021, 0.147]	0.144	55.4	40462
Sugarcane	0.082	[-0.02, 0.184]	0.117	29.9	19249
Sunflower	-0	[-0.065, 0.064]	0.995	40.9	28799
Wetland rice	0.035	[-0.08, 0.151]	0.550	48.7	35166
Wheat	0.058	[-0.018, 0.135]	0.135	36.0	24804
Normalized mean all crops	0.047	[-0.04, 0.133]	0.288	43.0	30498
Panel C: Amenities: Facility Available in Village? (1 = yes, 0 = no)					
Educational	0.001	[-0.045, 0.047]	0.971	62.2	45446
Medical	-0.059	[-0.129, 0.011]	0.101	49.1	35421
Drinking water	0.001	[-0.003, 0.004]	0.728	59.2	43274
Banking	0.026	[0.008, 0.044]	0.004	48.9	35277
Communication	0.015	[-0.048, 0.079]	0.641	58.5	42716

continued

Table 1: RD Estimates for Continuity of Covariates (Continued)

Postal	0.017	[-0.051, 0.084]	0.630	57.2	41791
Papers and magazines	0.032	[-0.049, 0.113]	0.438	52.1	37852
Panel D: Social and Demographic Characteristics					
Household size	0.027	[-0.082, 0.137]	0.626	41.6	29343
Literacy rate	0.008	[-0.013, 0.03]	0.454	39.9	28042
Log village area	-0.09	[-0.237, 0.056]	0.228	39.0	37060
Log village population	0.005	[-0.14, 0.151]	0.942	51.1	27258
Population share of SC/ST	-0.027	[-0.068, 0.014]	0.200	46.9	33757

Notes: Tests of continuity for covariates that are either fixed in time or unlikely to be affected by the presence of industrial pollution. Table reports geographic regression discontinuity estimates of the effects of severely-polluting industrial sites in villages immediately downstream of the sites. SC/ST refers to Scheduled Caste or Tribe, groups of historically marginalized people who are given special constitutional protections.

Table 2: Correlation of Satellite-based Proxies with District Agricultural Output

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

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

Table 3: RD Estimates for Crop Yield: Heterogeneity by Industry Type

Sample restriction	Estimate	Robust 95% CI	p-value	Bandwidth	Effective N
Panel A: Employment in polluting sectors within 25 km of site					
Below median	-0.015	[-0.072, 0.042]	0.605	39.3	8186
Above median	-0.033	[-0.089, 0.022]	0.236	62.4	31124
Panel B: Population within 25 km of site					
Below median	-0.024	[-0.064, 0.015]	0.227	48.9	14008
Above median	-0.036	[-0.103, 0.031]	0.290	60.8	26054
Panel C: Ratio of employment in polluting sectors to population					
Below median	-0.009	[-0.078, 0.06]	0.797	55.8	15884
Above median	-0.027	[-0.082, 0.027]	0.323	59.0	25416
Panel D: Downstream increase in Nitrate					
Below median	-0.078	[-0.203, 0.048]	0.224	39.3	4780
Above median	0.006	[-0.062, 0.074]	0.868	45.2	5172

Notes: Geographic regression discontinuity estimates of the effects of severely-polluting industrial sites on predicted crop yield in villages immediately downstream of the sites, for subsamples restricted based on characteristics of the industrial sites. We use employment in polluting industries (Panel A) as a proxy for the most severe pollution, population (Panel B) as a proxy for domestic and municipal water pollution, their ratio (Panel C) as a proxy for industrial pollution relative to domestic pollution, and the downstream increase in nitrate (Panel D) as a measure of potentially beneficial nutrients released at the site.

Table 4: RD Estimates for Other Surface Water Pollution

Dependent Variable	Estimate	Robust 95% CI	p-value	Bandwidth	Effective N
Nitrate (mg/L)	0.149	[-0.135, 0.432]	0.305	93.3	1016
Nitrite (mg/L)	0.108	[0.001, 0.215]	0.049	60.3	785
Fecal coliform	0.132	[-1.948, 2.212]	0.901	56.6	2591
Total Coliform	2.431	[0.451, 4.412]	0.016	57.3	2541
Calcium (mg/L)	195	[191.8, 198.2]	0.000	19.0	734
Magnesium (mg/L)	61.79	[59.68, 63.91]	0.000	31.2	1025
Sodium (mg/L)	408.3	[388.1, 428.5]	0.000	28.0	686
Chloride (mg/L)	384.4	[352.2, 416.6]	0.000	55.1	2230
Sulphate (mg/L)	284.7	[257.9, 311.5]	0.000	39.0	1360
Hardness (mg/L)	306.8	[286.6, 327.1]	0.000	24.6	928
Turbidity (NTU)	43.37	[40, 46.75]	0.000	25.1	915
Total Dissolved Solids (mg/L)	2181	[2041, 2321]	0.000	26.1	847
Total Fixed Solids (mg/L)	1743	[1661, 1825]	0.000	41.9	1081
Total Suspended Solids (mg/L)	80.16	[36.07, 124.3]	0.000	74.3	705
Conductivity < 1500	-0.367	[-0.367, -0.366]	0.000	13.5	614
pH in good range	-0.053	[-0.058, -0.047]	0.000	34.8	1534

Notes: Estimated effects of severely-polluting industrial sites on water pollution concentrations in nearby rivers, immediately downstream of the sites. Dependent variables are listed in rows. RD estimates as described in table 3. Sample includes villages within 20 km of a flow path that passes near each industrial site, as defined in the text. P-values are calculated using standard errors that are clustered by the monitoring station. NTU is Nephelometric Turbidity Units. Units for fecal and total coliform are 10^4 Colony Forming Units/mL. Units for conductivity are $\mu mhos/cm$.

Table 5: RD Estimates for District-level Actual Yield

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

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

4 References

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

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

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

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

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

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

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