The London School of Economics
and Political Science

Essays On the Economic Consequences of Environmental Pollution and Climate Change in India

Anshuman Tiwari

August 2022

A thesis submitted to the London School of Economics and Political Science for the degree of Doctor of Philosophy.

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work unless otherwise clearly indicated.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that this thesis consists of approximately 53,000 words.

Statement of Co-authored Work

One of the four chapters that form this thesis involves co-authored work:

Chapter 4 was co-authored with Nick Hagerty. Overall, my contribution amounts to 50 percent of the paper.

Abstract

This thesis seeks to enhance our understanding of how the natural environment interacts with the process of economic development in low- and middle-income countries. The context for study in this thesis is India, home to almost 18 percent of the world's population, and staring down climate change and various environmental pollution crises while pursuing economic growth to reduce widespread poverty. The four chapters uncover the *environmental* channels through which agricultural activity - the dominant source of employment in poor economies – affects, and is in turn affected by, the process of structural transformation. In chapter 1, I ask whether and how anti-poverty rural workfare programs affect the dependence of agricultural production on the weather. In chapter 2, I quantify the consequences of air pollution from seasonal but predictable agricultural fires on the size and spatial distribution of economic activity. In chapter 3, I document the economic costs from an increase in air pollution due to a groundwater conservation policy that unintentionally shifted agricultural fires into the winter. In chapter 4, I estimate the losses imposed by industrial water pollution on agricultural production. Taken together, these chapters drive home the vital importance of the natural environment to sustainable economic development.

Acknowledgments

I am enormously grateful to the various people who have supported the journey that has led to this thesis. It would not be possible without the invaluable supervision provided by Gharad Bryan, Ben Groom and Sefi Roth. The same holds true for Robin Burgess and Daniel Sturm, who were a source of constant support and guidance through various stages of my PhD.

I have received valuable advice, feedback and comments from many faculty members and staff at the London School of Economics and elsewhere: Gabriel Ahlfeldt, Nick Hagerty, Tim Besley, Swati Dhingra, Eugenie Duguoa, Marion Dumas, Roger Fouquet, Maitreesh Ghatak, Vernon Henderson, Lindsay Relihan, Lutz Sager, Cristián Sánchez, Misato Sato, Olmo Silva, Tom Smith and Chris Timmins. I would like to thank Philipp Barteska, Alix Bonargent, Shadi Farahzadi, Virginia Minni and Veronica Salazar-Restrepo for many discussions and for helping me stay sane.

I gratefully acknowledge funding from the Grantham Research Foundation, who also kindly hosted me within their offices at the LSE, as well as STICERD for a PhD research grant. Further thanks are due to administrative staff at the Department of Geography, Grantham Research Institute and the Department of Economics for their contributions in helping me reach this stage. I would also like to place on record my thanks to various seminar participants at LSE and many other places.

For their constant support, I am indebted to my family. For her enormous patience and unflinching confidence in me, for helping me stay grounded, and simply for being herself, I am eternally grateful to Ipsita.

Contents

List of Figures

List of Tables

Preface

There is a long tradition of research on how the *scarcity* of natural resources can constrain economic growth. More recently, the *degradation* of the natural environment in the form of climate change as well as air and water pollution has received increasing attention. This literature has mostly provided evidence on developed countries, whereas the costs of such environmental degradation may be much higher in Low- and Middle-Income countries (LMICs). These countries are characterized by widespread poverty, weak institutions, and low trust within society and in government, all factors that could potentially increase the damages from environmental degradation. In turn, the process of economic development itself can degrade the environment further in such settings (Jayachandran 2022). Therefore, the channel of causation can run from environmental degradation to economic development, or vice versa. In this thesis, I contribute to this comparatively nascent literature by documenting specific causal channels between economic development and environmental degradation in India.

India is home to almost 18% of the world's population who are expected to consume substantially more in the coming decades. Even as the process of structural transformation is underway, agriculture still is the main source of livelihoods for more than 50 percent of the population in the country. The incomes of cultivators and laborers reliant on this sector are especially vulnerable to climate change. At the same time, more than 82 percent of farms are operated by small and marginal farmers who practice subsistence agriculture and need constant government intervention such as workfare programs to supplement their meager incomes. Workfare programs can target consumption smoothing more efficiently and are also becoming increasingly common in response to extreme weather events driven by climate change. The first chapter investigates the effect of India's National Rural Employment Guarantee Scheme (NREGS) - the largest workfare program in the World - on agricultural productivity. I use the standard identification strategy in the NREGS literature that exploits program rollout across districts to document higher agricultural yield *losses* of 8 percent from negative rainfall shocks, after the program comes into existence. Next, I proceed to examine whether higher risk in crop choice is a potential mechanism that can explain this increased yield volatility. I construct novel crop risk indices using pre-NREGS moments of the crop revenue distribution, and confirm that higher risk as measured by these indices predicts higher realized yield after a positive weather shock, but lower yields after a negative shock. Using this measure of yield risk and the rollout strategy, I find little evidence that the increased yield volatility can be explained by higher risk in crop choice. On the contrary, and consistent with the literature, I find that NREGS strongly dampens pro-cyclical wage response to low rainfall shocks. This could increase labor costs and exacerbate the agricultural productivity effects of such shocks. Finally, higher provision of NREGS after a negative rainfall shock in a given year worsens yield losses the next year if a negative rainfall shock is also realized the next year, but improves yields if a positive shock is realized instead. Policymakers considering such programs should pay close attention to the negative and positive complementarities between social protection and agricultural productivity that these results suggest.

Historically, economic growth and structural transformation have tended to be accompanied by large-scale degradation of air and water quality, which may in turn affect economic growth. In the second chapter, I ask how consistently high levels of air pollution in some Indian cities affect aggregate productivity and spatial inequality across the country. I focus on air pollution from agricultural fires that are used to burn crop residue after the harvest. First, in order to quantify the effect of these fires on pollution both locally and in downwind districts, I build a novel econometric pollution transport model (ETM) and estimate its parameters using exogenous variation in yearly wind and cropping patterns. With this model, I demonstrate that external fires account for more than 10 percent of within-district annual variation in concentrations of particulate matter less than 2.5 microns in diameter (PM2.5) in Indian districts. I incorporate this ETM into a canonical quantitative spatial equilibrium model that features costly migration in response to both amenity and productivity differences across locations. Pollution affects both location amenity and productivity, the relative strengths of these mechanisms being governed by the pollution and income elasticities of migration respectively. Using the equilibrium equation describing migration shares across locations, I estimate these elasticites on migration shares data from the 2011 population census. I find that the pollution elasticity is not significantly different from zero although measured imprecisely, while the income elasticity of 0.81 is precisely estimated. I also estimate a parameter summarizing the institutional features that determine the prevalence of agricultural fires. Finally, I conduct model counterfactuals to reduce fires in the worst-offending states of India and find an increase in national GDP of 1.22 percent and a reduction in the Gini coefficient of GDP by 0.23 percent, since large reductions in pollution in the poorest Northern parts of India leads to a reallocation of labor towards the North.

In the third chapter, I describe how fixing one environmental externality in the second-best setting of LMICs can exacerbate another environmental externality. Alarming rates of groundwater aquifer depletion in North India are linked to water-intensive rice cultivation based on cheap electricity for water pumps. Since optimal marginal pricing of groundwater is not politically feasible, the northwestern states of Punjab and Haryana that have high groundwater depletion rates, instead passed legislation in 2009 with the intent to rely more on rain-fed irrigation by mandating a delay in rice crop transplantation to coincide with monsoon arrival. At the same time, rice crop residue burning in these two states contributes to high PM2.5 levels over North India. I use satellite data on fires and a difference-in-differences framework to document that the groundwater laws shifted more than half of all agricultural fires into early winter, when meteorological conditions favor longer suspension of particulate matter over North India. I then quantify the consequences of this increased air pollution on Indian GDP by estimating two further elasticities. First, I develop a novel instrument for PM2.5 that summarizes the exposure of a given location to upwind fires, showing that a 10 percent higher exposure to November fires increases annual PM2.5 concentrations by 0.3 percent in the average district. Second, I estimate the effect of higher PM2.5 concentrations on GDP with new data on Indian districts between 2007-2013, using district and year fixed effects combined with a first differences approach that is more efficient for non-stationary data, and with the fire exposure instrument to tackle residual reverse causality. With this approach, I find estimates that a 10 percent increase in PM2.5 reduces GDP by 1.8 percent on average, with a 95 percent interval of [-0.4%, -3.17%]. With these two elasticities and the structure of the instrument, I estimate that the groundwater laws decrease yearly Indian GDP by 0.125 percent due to the increase in November fire-driven air pollution.

In the co-authored fourth chapter, we shift our attention to the impact of industrial water pollution on agricultural output in India. This type of pollution, a by-product of industrialization and structural transformation, is high in many developing countries, but researchers and regulators have paid it less attention than air and domestic water pollution. We focus on 71 industrial sites identified by the central government as "severely polluted." We exploit the spatial discontinuity in pollution concentrations that these sites generate along a river. First, we show that these sites do in fact coincide with a large, discontinuous rise in pollutant concentrations in the nearest river. Then, we find that remote sensing measures of crop growth are 2.6 percent lower in villages downstream of polluting sites, relative to villages immediately upstream of the same site in the same year. In terms of agricultural production, this estimate roughly translates to a 1 percent decline in crop yields. The effect appears to be driven by reduced yields per cropped land area and not factor reallocation. These results suggest that damages to agriculture may not represent a major cost of water pollution, though many other potential social costs remain unquantified.

Chapter 1

Do Workfare Programs Affect Agricultural Risk through Crop Choice?

1.1 Introduction

Agriculture is the largest source of livelihoods in most low and middle-Income Countries (LMICs). Weather risk is a pervasive feature of agriculture in these settings, with the effects of climate change set to make this risk worse in the future decades (Hallegatte et al. 2016). The lack of insurance against such weather risks leads to substantial welfare losses (Dercon 2002), reduces agricultural productivity (Cole and Xiong 2017) and inhibits productive investments (Morduch 1995). How would indemnifying income risk from weather shocks affect aggregate agricultural output? This paper analyzes the effect of large-scale workfare programs, which reduce income risk for vulnerable populations, on aggregate agricultural yields, and investigates whether risk in crop choice can explain the observed effect.

Workfare programs provide livelihood support to the poorest, particularly after income losses from events such as adverse weather shocks. Their self-targeting mechanism makes these program attractive in the absence of unemployment insurance in fiscally-constrained LMICs (Ravallion 1991; Besley and Coate 1992; Bertrand et al. 2021). Such programs are also increasingly being considered part of a flexible climate adaptation strategy (Rigolini 2021), given that private adaptation may be limited (Burke and Emerick 2016; Taraz 2017, 2018; Fishman 2018). The National Rural Employment Guarantee Scheme (NREGS) in India is the largest such program in the world, promising at least 100 days of minimum wage manual work to each household that demands it.

This paper documents the effects of NREGS on the volatility of agricultural yields. I combine the standard identification strategy in the NREGS literature that relies on the rollout of the program across Indian districts with exogenous yearly weather shocks, controlling for time trends and also making use of a first difference specification. The program decreases aggregate yields by an additional 10% after a negative rainfall shock, the same magnitude of yield loss to a similar rainfall shock pre-program. This effect is precisely estimated and is consistent across various specifications including the first differences specification that performs better for strongly non-stationary data (Wooldridge 2010).

Next, I investigate whether aggregate risk in crop choice is a potential mechanism that could explain the additional volatility. Most small and medium-sized farm-households provide some labor to the agricultural labor market in India, apart from cultivating their own fields; the smallest farm-households are net sellers of labor.[1] The NREGS literature documents substantial increases in prevailing wages for manual farm labor, using both natural experiments (Imbert and Papp 2015; Berg et al. 2018) and RCTs (Muralidharan et al. 2016). This general equilibrium increase in agricultural wages driven by NREGS increases expected earnings for these small farm-households. Through the consumption smoothing opportunities available in the event of agricultural productivity shocks, NREGS can act like social insurance. The provision of such insurance may therefore induce small and medium farm-households to take on additional agricultural risk. Given that most farm-households are smallholders in India, this increase in individual risk could affect aggregate risk.

[1] The median farm size in India is extremely small at 0.9 acres, leaving substantial family labor available for hire on the agricultural market.

In order to test this mechanism, I construct novel indices of aggregate risk using pre-NREGS moments of the district crop revenue distribution. District-crop area shares for later years are used to aggregate each of these three moments into three separate indices. Yearly variation in these indices comes from changes in the district-level crop mix. For example, a relative increase in area under crops with higher standard deviation of pre-NREGS revenue would increase the Risk Index of Crop Choice constructed using the second moment (RICC-SD). I show that these indices have skill in predicting variation in realized crop revenue. For example, higher risk in crop mix as measured by RICC-SD is positively correlated with realized crop revenue in normal weather years, but decreases realized revenue in years with bad rainfall or higher than normal temperatures, and increases realized revenue during a good rainfall year. These findings build confidence that these risk indices are meaningful measures of aggregate crop choice.

Using the standard NREGS identification strategy reliant on rollout across districts, I find no evidence of changes in aggregate risk as measured by risk indices constructed from the first and third moments of pre-program crop revenue. I find that the risk index constructed using the second moment increases very slightly by 0.08% after NREGA, indicating a minuscule shift in cropped area toward more risky crops. But, this increase can only explain less than 1% of the net additional effect of NREGA on the rainfall sensitivity of crop yields. Therefore, aggregate risk in crop choice, as measured by the risk indices I construct, does not seem to be driving the increased rainfall sensitivity.

The labor market channel might help explain the effects. Imbert and Papp (2015) find average wage increases of about 8% due to NREGA while Muralidharan et al. (2016) find that beneficiary households' earnings increase by about 14%. These are large effects of a similar magnitude to the estimates for increased sensitivity of crop yields. I further confirm findings in (Rosenzweig and Udry 2014; Santangelo 2019) that wages become less elastic (by about 5.5%) to rainfall shocks after NREGA. This inability to modulate wages in a pro-cyclical manner after a bad rainfall shock may increase labor costs enough for some larger farm-households that it could hurt output during harvest; it may also cause some smaller farm-households to abandon their own crop in order to earn higher incomes on the private labor market.

Two other channels through which NREGS might affect agricultural outcomes are the provision of community infrastructure such as irrigation through public works, and higher use of inputs such as fertilizer due to an alleviation of credit constraints. Neither of these channels would explain why crop yields worsen with negative rain shocks after NREGS; better irrigation would make yields *less* sensitive to rainfall shocks while higher fertilizer usage would increase expected yields without affecting volatility.

This paper contributes to the limited literature on the impact of workfare programs on agricultural outcomes. Firstly, in parallel with (Taraz 2021), this paper documents increased rainfall sensitivity of crop yields post NREGS. This paper builds further confidence in the increased sensitivity result by using two years of additional data as well as a first difference specification that deals better with non-stationary data. But this paper also explicitly analyzes aggregate risk in crop choice as a potential mechanism, finding that it is unlikely to explain the increased crop yield sensitivity. Secondly, I use a nationwide data set of total agricultural output relative to the existing literature on the impact of NREGS on agriculture which focuses on data from one state (Gehrke 2019) or a representative sample rather than complete population from administrative data (Deininger et al. 2016).

Thirdly, I contribute to the small literature on crop choice and climate change by constructing a novel measure of aggregate risk in crop choice. Most papers in this literature use discrete choice models to understand the determinants of cropping patterns (Seo and Mendelsohn 2008; Wang et al. 2010; Kurukulasuriya and Mendelsohn 2008). An exception is Auffhammer and Carleton (2018) who study the impact of crop diversity on drought resilience. In contrast to the literature using discrete choice models, I utilize OLS regressions with a transparent identification strategy. I also study crop choice as an optimal risk-taking response to NREGS-as-insurance that might cause increased yield volatility, relative to other literature which studies crop choice as a climate adaptation margin.

The rest of the paper is structured as follows: section 2 describes the workfare program, section 3 surveys related NREGS literature, section 4 discusses a theoretical framework, section 5 describes data sources and construction of outcomes, section 6 relates the research design, section 7 presents the results and section 8 concludes.

1.2 Background

1.2.1 The workfare program

The NREGS program was created through an Act of Parliament in 2005 that provided a legal right to employment on labor-intensive public works on demand to each rural household for a minimum of 100 days. The key feature of the program is that it provides a minimum of 100 days of employment per household on demand on public works activities. It incorporates labor-intensive minimum wage work requirements such that individuals with a high opportunity cost of time select out (Besley and Coate 1992). The local administration is supposed to provide work within 5 km of home and 15 days of application.

The program was introduced in phase I to the poorest 200 districts in February 2006, followed by the next poorest 130 districts in phase II (February 2007) and the remaining districts in phase III (April 2008). The assignment of district to phase was based on a "Backwardness Index" created by the Planning Commission using data from the early 1990s. Variables that were used to determine this index along with the weights are available online[2]. The actual assignment of districts to phases

[2] The variables include the fraction of lower castes (constitutionally protected underprivileged groups), agricultural productivity per capita and log casual agricultural wage respectively

did not perfectly follow the index since there was a lot of political bargaining over the large budget allocation to the program. For instance, each state had to have one district in each phase, regardless of the rank of the district. Hence some poor districts in rich states got the program over a poorer district which is among the richest ones in a poor state.

In 2010-2011, 2.3 billion person-days of employment was generated among 53 million households. The budget for that year was Rs 345 billion (US$1.64 billion, 0.6 % of GDP). 60% budget of the total budget is supposed to be for wages and 33% of work is reserved for women at an equal wage to men. Among the projects to be undertaken as part of the program, water management is a major goal. This includes micro-irrigation works, drought-proofing and flood-proofing. The local village council approves projects in consultation with block and district administrations.

The program comes with exhaustive and detailed operational guidelines that run to over 200 pages[3]. This does not preclude further ad-hoc documents that govern aspects of the program separately. A common finding in the literature on NREGS within economics, political science and other related social sciences is the heterogeneity in implementation of the program. Some reasons for this in the literature include the varying nature of labor market conditions and need for public employment, differing administrative and fiscal capacities of states, local elite control and politician-bureaucrat dynamics (Sukhtankar 2017).

1.2.2 Related literature

Evidence on the impact of NREGS on agricultural output and yields are thin compared to the evidence on labor market, consumption, education and other development outcomes. The few articles on this topic are reviewed next. Gehrke (2019) uses panel data from the Young Lives study in Andhra Pradesh to show that house-

[3] https://NREGS.nic.in/Circular_Archive/archive/Operational_guidelines_4thEdition_eng_2013.pdf

holds use more inputs on cotton (a commercial crop which is more risky than food grains such as rice) after the introduction of NREGS. Deininger et al. (2016) use household panel data from the Additional Rural Incomes Survey and Rural Economic and Demographic Survey (ARIS/REDS) to similarly show that area devoted to rice goes down even as area under high value crops go up. They also show that percentage irrigated area and input usage increase along with the number of crops planted in all cropping seasons. These papers make the argument that the implicit insurance provision in NREGS allows small farmers to diversify crop portfolios by growing more risky crops and also increases the number of crops being planted in all the cropping seasons. Santangelo (2019) utilize nationally representative employment data to show that the relationship between rainfall shocks and agricultural yield does not change after introduction of the NREGS but that rural wages are no longer pro-cyclical, i.e., NREGS weakens the impact of rainfall shocks on local rural wages. Further, Bhargava (2013) uses agricultural census data to show that smaller farmers are more likely to adopt mechanical technologies in response to rising wages.

In the paper that is closest to my study, Taraz (2021) documents increased volatility of aggregate yields to rainfall shocks after NREGS comes into force, using the same district agricultural output panel data set that I employ. Her findings for the increased sensitivity are of a similar magnitude to those found in this paper. In contrast to Taraz (2021), I extend the analysis to include two additional years of data to 2013. In contrast to their usage of a standardized precipitation variable for rainfall shocks, I utilize the definition of rainfall shock as a dummy based on deviations from historical records (uses the) that has been widely used in the literature on wage determination in Indian village economies (Jayachandran 2006; Kaur 2019). While results are similar in both cases, the use of a dummy allows easier comparison with the previous literature. More importantly perhaps, Taraz (2021) does not test any potential mechanism, only suggesting that various possible mechanisms could explain the result. I develop novel measures of risk in crop choice and proceed to show that increased risk, as measured by these indices, does not explain much of the increased crop volatility.

While the consensus in the literature is that NREGS has increased rural wages, whether this is due to productivity increases or increased market competition for labor is unexplored (Sukhtankar 2017). There is also considerable evidence that NREGS may have crowded out private labor supply (Azam 2012; Imbert and Papp 2015; Berg et al. 2018; Muralidharan et al. 2016). There is some disagreement about which parts of private work declines - most of the evidence suggests that the fall in private sector work may represent a fall in disguised unemployment, idle time or private work with close to zero productivity but Deininger et al. (2016) find that on-farm self-employment increases. All but one of these papers utilize a difference-in-differences strategy that arises from a phased rollout of the program but use various data sources. Imbert and Papp (2015) and Azam (2012) use the National Sample Survey data, Berg et al. (2018) use the Agricultural Wages data from the Indian Ministry of Agriculture while Deininger et al. (2016) use the ARIS/REDS household panel data. Muralidharan et al. (2016) run an RCT in the state of Andhra Pradesh that evaluates the impact of a reform of the NREGS delivery system in the state of Andhra Pradesh. Hence they are able to collect their own data with experimental variation in the improvement of NREGS implementation.

As mentioned earlier, there exists a vast literature on the impact of NREGS on other development outcomes. The reader is referred to the excellent survey by Sukhtankar (2017) for an exhaustive appraisal.

1.3 Theoretical Framework

This paper investigates whether NREGS makes crop yields more sensitive to weather shocks, and whether this finding can be explained by increased aggregate risk in the district crop mix. I now provide a theoretical treatment of these questions below and discuss the resulting testable hypotheses.

1.3.1 Risk in crop choice

Risk in production decisions is an important characteristic of the environment for rural farmers in developing countries.[4] These farmers are usually characterized as risk-averse given that they are extremely poor and lack reliable consumption smoothing in the event of productivity shocks. Such risk aversion may cause farmers to plant lower yielding crops that also carry lower output risk. Insurance for output risk could enable such risk-averse households to make more optimal crop choice decisions; but such insurance is usually not available. This insurance market failure can lead to the perpetuation of a low productivity equilibrium (Cole and Xiong 2017; Morduch 1995).

The bulk of agriculture in India is carried out by small farm-households; the median household farm size is about 0.9 acres (Kaur 2019). These households are more likely to be net sellers on the agricultural labor market (Imbert and Papp 2015). NREGS increases the net incomes of such farmers by providing work during the lean season and in the aftermath of a poor monsoon. This provision of work at close to or higher than agricultural wages introduces an entirely new consumption smoothing mechanism. Thus NREGS can be interpreted as a public insurance program that can more than supplements net incomes when agricultural productivity is low. Muralidharan et al. (2016) show, in an RCT, in the state of Andhra Pradesh that average earnings of a rural household increase by 14%, with 2/3rd of the gains coming from the general equilibrium wage increases which benefit small farm-households the most. These are large effects that could raise expected incomes substantially.

Viewed from the lens of portfolio choice theory, this provision of insurance to risk-averse farmers could lead them to take on higher risk in their crop choice portfolio. Higher risk may be individually optimal in this setting given the insurance market

[4] This includes both output and price risk - in this paper, I consider output risk only since price-fixing mechanisms such as government Minimum Support Prices (MSP) are a common feature of this setting, in theory limiting the price risk faced by farmers.

failure that forces the choice of lower risk portfolio in the first place, and could increase expected aggregate yields within district. But higher risk could increase the volatility of yields by making aggregate yields more sensitive to adverse weather shocks.

1.3.2 Other mechanisms

A few other mechanisms have been postulated in the literature for how NREGS could affect aggregate yields and the weather sensitivity of yields. This paper limits itself to testing the Crop Choice mechanism. But I provide a short summary of these other mechanisms below.

1.3.2.1 Labor market channel

Before NREGS, weather-driven negative productivity shocks would not change market labor supply or would even *increase* it, since labor was the only economic resource small farmers could sell to smooth consumption (Jayachandran 2006). Larger landowners may have benefited from the pro-cyclical downward wage adjustment that occurs when labor demand decreases while labor supply does not.

An important general equilibrium effect of NREGS is the increase in agricultural wages documented in the literature. At the same time, agricultural wages may become less elastic to negative productivity shocks after NREGS (perhaps because the program creates an outside option that may reduce the monopsony power of village landowners) (Santangelo 2019). Nominal wage rigidities in Indian village labor markets documented by Kaur (2019) can further solidify any level increases in the agricultural wage due to NREGS, and reduce counter-cyclical wage responses to productivity shocks.

This increase in the wage level may negatively affect crop yields if farmers who are net buyers of labor cannot afford to hire more expensive labor during harvest, or if availability of labor is constrained. Farmers that are net sellers of labor may shift labor supply away from own-farms if their outside earnings are higher than the shadow wage on their own farm. Secondly, yield losses from a weather shock could be exacerbated after NREGS if wages do not adjust downward, thereby further increasing labor costs and decreasing labor availability. In the long run, some farmers may adjust to higher costs by increasing mechanization or diversification into non-agricultural activities, avenues that are usually not available to smaller farmers (Bhargava 2013).

I provide some corroborating evidence on the labor market impact of NREGS that confirms existing findings of an increase in wages. The labor market mechanism would result in lower expected yield while making yields more sensitive to adverse weather shocks.

1.3.2.2 Insurance channel

While the effects of NREGS-as-insurance on crop choice are detailed above, this channel could also affect other agricultural practices such as the adoption of more resilient seeds or better inputs. While mechanization can be seen as a response to increasing wages, the procurement of high fixed cost machinery could also be enabled by the higher incomes that small and medium farmers earn from NREGS. These mechanisms would increase expected yield but also make yields *less* sensitive to weather shocks

1.3.2.3 Infrastructure channel

Decisions on the public works programs to be undertaken under NREGS are supposed to decided by the local community; in practice this is rarely the case, with a top-down approach more common (Khera 2011). While corruption and misuse of funds under NREGS were documented in the earlier years (Dutta et al. 2012), administrative reforms including better monitoring mechanisms through the use of MIS systems were gradually instituted. If such public works lead to better community-level provision of productive infrastructure such as irrigation facilities or flood protection mechanisms, the level of yields would increase while decreasing yield sensitivity. These mechanisms would also reduce sensitivity of yields to weather shocks while increasing expected yield.

1.4 Data

1.4.1 Agricultural outcomes

While the ideal outcome measure to use would be farm-level profits for each crop over time, such data are seldom available for any country. Therefore, I rely on aggregate measures at the district level in India compiled by the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in their District Level Database (DLD).[5] This data contains information on crop area planted, output and prices for all the main crops as well as some peripheral crops. Price data is available for 16 crops, covering about 79% of all area under cultivation. This data contains 571 districts across 20 states from 1990-2015 for the agricultural year that runs from July 1 to June 30. I describe the construction of the outcome measures to capture aggregate crop yield and risk in crop choice below.

1.4.1.1 Measure of crop yields

The main measure is a crop area-weighted sum of the revenue value of output per hectare for each crop ("Revenue Value of Yield" or RVY). Since price data is patchy, I construct single national prices for each crop from pre-program data. All price data used in the analysis pertains to these single national prices. This measure of crop yields captures aggregate crop yield in a single index. This measure has been used widely in the literature to capture output losses without being affected by changes in prices (Duflo and Pande 2007; Burgess et al. 2017; Taraz 2021).

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

Indian agricultural markets are heavily regulated including through the Minimum Support Price (a floor on crop prices); these markets are also typically not well-integrated across district. These forces can cause prices to move in opposite directions, with trade frictions compensating farmers for some of the crop output losses through higher prices. But the Revenue Value of Yield captures the output loss that is the focus of this paper, leaving price effects out. I also make use of a crop area-weighted output per hectare as a second measure that does not utilize price data.

1.4.1.2 Measure of risk in crop choice

Output risk is a major issue for farmers in response to negative productivity shocks; higher prices can only compensate for part of the losses due to such shocks. To capture this output risk, I calculate the first three moments of Revenue Value of Yield (RVY) for each crop using data from before 2003. Each of the three measures of risk index of crop choice for each district-year from 2003 onward are then constructed through weighted sums of these three moments for each crop, with the weight being the yearly share of area planted under each crop in that district. The distribution of pre-program RVY is calculated per crop in a contiguous region that shares similar agricultural characteristics, striking a balance between sample size and variation across districts. These moments capture relative crop risk within each region.[6]

The volatility of an asset is usually measured with the second moment of its distribution. The Risk Index of Crop Choice constructed using the second moment ("RICC-SD") can be interpreted as measure of expected volatility of revenue from the district crop mix. Assets with higher returns also tend to be more volatile; this is also true of crop revenues. For this reason, the RICC-Mean and RICC-SD are strongly correlated. I also study whether farmers switch to crops with more out-

[6] There are an average of 4.8 districts per region. I also conduct robustness using moments from each crop-state combination. The results are presented in the appendix.

liers using the skewness of pre-program RVY. A positively skewed distribution tends to have more positive outliers than a normal distribution. In other words, returns from these crops are likely to be relatively higher with good rainfall. The skewness of assets is a major consideration in the financial economics literature. For example, Mitton and Vorkink (2007) find that underdiversified investors have a preference for positively skewed stocks; Barberis and Huang (2008) show that a positively skewed security can be "overpriced" and can earn a negative average excess return; and Zhang (2018) demonstrates that Swedish retail investors with lower wealth or labor incomes that have higher downside risk tend to seek investment portfolios with higher skewness

An important point to note is that such changes reflect only yield variation and not price variation, since single national prices are used to construct pre-program moments of the Revenue Value of Yield across regions. Since the moments for each district-crop are fixed at pre-program levels, yearly changes in RICC comes from changes in the area planted under various crops.

Figure 1.1 plots the variation in the revenue value of yield and the three measures of risk in crop choice over the sample period, separately for the three NREGS phase districts. Only the revenue value has a growth trend across the sample period.

Figure 1.2 plots the mean for the distribution of crop-region revenue value of yield between 1990-2002 against the SD of this distribution. The figure reflects the fact that a more volatile crop that carries higher risk also has higher reward, although it also comes at a higher cost of inputs.

1.4.2 Wages

This paper does not extensively test for the labor market channel. However, I do provide corroborating evidence to the wage increase with NREGS as well as the reduced weather sensitivity of wages to productivity shocks post-NREGS. The data used is described in this section.

The National Sample Survey on Employment and Unemployment (NSS EUE) is the main source of information on labor conditions including wages and employment in India. I make use of the NSS EUE rounds from years 2003, 2004, 2005, 2007, 2009 and 2011 respectively. This survey provides a detailed break-up of the time spent by activity status of each individual in the survey for the 7 days prior to the survey date. The survey covers every member of the household, regardless of age.

I follow Imbert and Papp (2015) in limiting the sample to individuals aged 18-59 to construct the wage data. Public works employment on NREGS is the closest substitute to "casual labor," which is usually work done on a daily wage rate on spot labor markets. The NSS EUE differentiates such casual work with a separate activity status. I create casual labor wages using this activity status, constructing daily rates for each individual between 18-59 years of age.

1.4.3 Weather data

1.4.3.1 Heating degree days

I calculate heating degree days (HDD) for each district-year separately for the planting season and growing season using the ERA5 reanalysis dataset from the European Center for Medium Range Weather Forecasting (ECMWF).[7] This measure of heat exposure is a metric proposed in the agronomic literature, and has been commonly used to capture crop output losses from total excess heat exposure on each growing plant organism (D'Agostino and Schlenker 2016; Burke and Emerick 2016; Colmer 2021). The basic idea here is that temperatures up to the threshold might not hurt the organism or might even be beneficial, but above the threshold the organism suffers harm that is captured well through a linear approximation in total temperature exposure above the threshold. The HDD measures the number of heating degree-days above a threshold during a particular period of

[7] https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

time.

$$HDD_{Threshold}(T) = \sum_{season} (T - Threshold) * \mathbb{1}(T > Threshold)$$

ERA5 provides hourly data on temperature at 30 km resolution for the whole world from 1979 onward. I follow the literature in using the sine interpolation method to calculate the fraction of each day above the threshold temperature (D'Agostino and Schlenker 2016). To calculate total heating degree days, I sum up the total excess temperature over the period under consideration. Finally, I calculate district-level HDD using inverse square distance weighting from the district centroid.

I calculate HDD for various thresholds and seasons. Colmer (2021) shows that crops differ in their optimal HDD threshold values; I assign the threshold for individual crop yields based on their calculations. A single threshold is necessary for aggregate crop yields; I choose 25 C as the main threshold and conduct robustness for other thresholds. The main growing season months for most of India are the main monsoon ("Kharif") season of June-October, and I consider those months in the main HDD calculation.[8]

1.4.3.2 Rainfall shocks

To capture the effect of precipitation on aggregate crop yields, I follow Jayachandran (2006) and Kaur (2019) in constructing a piecewise function of rainfall. First, I calculate total precipitation in the planting and growing seasons for each district-year using daily rainfall data from TerraClimate provided by the Climatology lab.[9] Next, I calculate the twentieth and eightieth percentiles of each district's historical precipitation record. Then I designate rainfall extremes by creating high and

[8] Kharif season Crops in some districts follow a different calendar, whereas Winter ("Rabi") season runs from roughly November to March. However, Colmer shows that monsoon season rainfall and temperature are important even for Rabi crops, since the amount of moisture retained in the soil through to the Rabi season depends on temperatures in the monsoon season. I do robustness around the season considered for the weather variables in the appendix

[9] https://www.climatologylab.org/terraclimate.html

low rainfall indicators as follows. The high rainfall dummy (*high_rain*) turns on if precipitation is above the eightieth percentile for the district, and zero otherwise; similarly the low rainfall dummy (*low_rain*) turns on if precipitation is lower than the twentieth percentile of historical precipitation, and zero otherwise. This non-linear function has been extensively used to capture the effect of rainfall shocks on agricultural productivity, especially in India and allows us to flexibly capture the effect of excess and deficient rainfall on aggregate crop yields.

1.4.4 Further district controls

I construct district level controls from the Indian National Census 2001: fraction population that is SC/ST (caste groups that have historically been discriminated against), population density, literacy rate, male and female labor force participation ratio, fraction of labor force in agriculture, irrigated cultivable land per capita and non-irrigated cultivable land per capita. I also use the NSS and crop data to create controls for baseline agricultural wages and agricultural productivity per worker.

1.4.5 NREGS data

The NREGS program was rolled out over a period of three years across the whole of India, as detailed previously. This information is available on the website of the ministry of rural development.[10] I use three district-level NREGS take-up measures: the number of NREGS person-days worked, the number of households working the maximum number of days permitted, and, NREGS labor expenditure. The NREGS data corresponds to the fiscal year (April 1 to March 31) and is available for 2006–2012. Administrative reforms in 2008 reduced large-scale corruption issues from inflated reporting in the official NREGS reports relative to survey data (Imbert and Papp 2015).

The provision of NREGS also fell dramatically in 2014 with the election of a new

[10] Can be found at https://NREGS.nic.in/MNREGS_Dist.pdf

central government which was opposed to the program. Therefore, I limit the analysis to the agricultural year 2013-14.

1.4.6 Indian district administrative boundaries

In order to calculate weather data at the district level, I make use of publicly available district administrative boundaries in the form of shapefiles from the 2011 census[11].

1.4.7 Construction of district panel

There were 593 districts in the 2001 Indian census. Many district administrative boundaries changed over the 2001-2011 period, due to creation of new states or to provide better administrative efficiency through smaller districts. In order to construct a panel of districts over the 2001-2007 period, I began with a list of unchanged census districts from 2001 and 2011[12]. This master list is then sequentially matched with the NSS districts, crop data districts, the administrative boundary districts and the NREGS districts.

At the end of this process, I am left with 466 districts that form a panel from 2003-2013. I start the empirical analysis in 2003 because I use data from 1990-2002 to construct data on the variables used to construct the NREGS backwardness index is not available for districts in the panel before this year. I end the analysis in 2013 since the new government drastically reduced provision of NREGS after taking office in 2014, in keeping with their electoral promises.

Table 1.1 provides summary statistics for the baseline covariates used to assign NREGS districts, the main outcome and explanatory variables as well as controls. The table disaggregates this information by the three NREGS phases.

[11] I make use of the shapefiles provided by the Datameet google group

[12] Available at http://censusindia.gov.in/2011census/maps/administrative_maps/Final%20Atlas%20India%202011.pdf

1.5 Research Design

1.5.1 Effect of weather shocks on crop yields

Weather shocks are an important predictor of crop yields in the literature. I document the importance of these weather shocks for individual crops in this section. I run the following regressions.

$$Weather_{dy} = \{HDD_{dy}, low\_rain_{dy}, high\_rain_{dy}\}$$

$$crop\_yield_{dy} = \tilde{\delta}\, Weather_{dy} + D_d + Y_y + \epsilon_{dy} \tag{1.1}$$

The vector $\tilde{\delta}$ denotes the effect of weather shocks on crop yields. The coefficient δ^{hdd} on HDD captures the average effect of an extra degree-day over historical levels on crop yields. Similarly, the coefficient on low_rain ($high\_rain$) captures the average effect of a low rainfall shock on crop yields. I use the data from 1990-2013, including pre-program years to maximize power. The HDD thresholds and growing seasons are taken from Colmer (2021), who choose these by maximizing R-squared from various regressions with different seasons and HDD thresholds. In the rest of the paper, for aggregate crop yields, the HDD threshold is 25 and the growing season is June-October.

Identification relies on the exogeneity of weather shocks controlling for district and year fixed effects (D_d and Y_y respectively). This assumption is quite common in the literature and does not raise any concerns in this setting either, given that yearly weather shocks are as good as randomly assigned. The main concern with inference is the spatial correlation in these shocks; I calculate Conley standard errors that account for this spatial correlation and also for autocorrelation across an arbitrary number of time periods using R routines provided by Thiemo Fetzer.[13] This approach toward inference is continued in the rest of the paper.

1.5.2 Effect of weather shocks on NREGS provision

Before conducting the main analysis, I test whether the provision of NREGS responds to adverse weather shocks. Evidence from studies such as Dutta et al. (2012) points to large demand for public works not always being met due to rationing. In this sense, administrative data on the quantum of money spent on labor, person-days worked or number of households that worked over 100 days are a result both of the demand for public works but also supply constraints by bureaucrats.

Therefore administrative data do not allow us to parse out whether labor demand on NREGS is higher during adverse weather shocks; rather, they allow us to test whether these measures - which depends on both demand and supply for public works - respond to adverse weather shocks. Given the corruption issues in the early years of NREGS implementation that were corrected by administrative reforms in 2008, I restrict these regression to 2009–2013, since these data come from the administrative system.

[13] See http://www.trfetzer.com/using-r-to-estimate-spatial-hac-errors-per-conley/

$$NREGS\_admin\_measure_{dy} = \tilde{\kappa}\, Weather_{dy} + D_d + Y_y + \epsilon_{dy} \qquad (1.2)$$

The vector $\tilde{\kappa}$ denotes the effect of weather shocks on NREGS provision measures. Identification relies on the exogeneity of weather shocks controlling for district and year fixed effects (D_d and Y_y respectively). This assumption is quite common in the literature and does not raise any concerns in this setting either, given that yearly weather shocks are as good as randomly assigned. The main concern with inference again is the spatial correlation in these shocks which I tackle with the same approach described in previous section.

1.5.3 Effect of NREGS on weather sensitivity of crop yields

Equation 1.3 presents the regressions I run to test for increased weather sensitivity post NREGS. The outcome variable is the Revenue Value of Yield, the main measure of aggregate yields.[14]. The coefficients of interest is $\tilde{\beta_1} = \{\beta_1^{hdd}, \beta_1^{lrain}, \beta_1^{hrain}\}$; if these are different from zero then the sensitivity of aggregate yields to weather shocks is different post-program relative to pre-program sensitivities $\tilde{\gamma_1} = \{\gamma_1^{hdd}, \gamma_1^{lrain}, \gamma_1^{hrain}\}$. Identification of $\tilde{\beta_1}$ (and $\tilde{\gamma_1}$) relies on the exogeneity of yearly weather shocks. I use data from 2003-2013 since data from before 2003 are utilized to estimate the main measure of risk (area-weighted SD of the pre-program revenue value of yield).

[14] I also plan to present a measure that does away with price data

The main threat to identification for $\tilde{\beta_1}$ is that another variable modulates the effect of weather shocks on crop yields at the same time as the program rolls out. In particular, differential time trends across the poorest districts which were targeted first by the program could be an issue. In order to alleviate such concerns, I run various specifications that account for this potential issues; equation 1.3 presents the most saturated specification using fixed effects.

In the first specification, I include the NREGS program dummy, weather variables as well as interactions of the three weather variables with the NREGS dummy. In the second specification, I allow a linear time trend interacted with a phase dummy. This controls for differential time trends in the outcome that differ by NREGS phase. In the third specification, I allow for these time trends to differ for each district based on initial values of observable characteristics that were used in the allocation of districts to NREGS phase. Fourthly, I interact weather shocks with values of these initial values of these characteristics to allow them to mediate the effect of weather shocks separately for each district.

$$log(RVY_{dy}) = \alpha_1\, NREGS_{dy} + \tilde{\gamma_1}\, Weather_{dy} + \tilde{\beta_1}\, NREGS_{dy} * Weather_{dy} + \\ \lambda_d^p * t + \phi_1^1\, Z_d * t + \phi_1^2\, Weather_{dy} * Z_d + D_d + Y_y + \epsilon_{dy} \tag{1.3}$$

RVY is the revenue value of yield, $NREGS$ is the program dummy, $Weather$ is a vector containing $\{HDD$, low_rain and $high\_rain\}$, Z_d is a vector containing pre-program values of the variables entering the "backwardness" index, λ_d^p denotes the NREGS phase the district was part of, and D_d and Y_y are district and year fixed effects respectively.

While controlling for trends in these specifications allows us to build more confidence in the results, I also report results using first differences. First differences (FD) can make non-stationary data stationary and be more robust than fixed effects (FE) when data have strong autocorrelation, as can be seen in panel (a) of figure 1.1. Further, an FD specification that also includes a fixed effect allows for a district-specific linear growth rate g_d in the outcome. The FD approach is commonly used in the macroeconomic literature to deal with serial correlation in aggregated GDP data, similar to the measures I use in this paper. Equation 1.4 specifies the regression framework for the FD model.

$$\begin{aligned}\Delta log(RVY_{dy}) = \alpha_1\, \Delta NREGS_{dy} + \tilde{\gamma_1}\, \Delta Weather_{dy} + \\ \tilde{\beta_1}\, \Delta(NREGS_{dy} * Weather_{dy}) + g_d + \Delta Y_y + \Delta\epsilon_{dy}\end{aligned} \tag{1.4}$$

In both the panel and FD specifications, the coefficient α_1 captures the average effect of NREGS on revenue value of yield during normal weather years. A large NREGS literature uses similar difference-in-differences design with twoway fixed effects (TWFE) for district and year to estimate average effects on various outcomes (Imbert and Papp 2015; Berg et al. 2018; Gehrke 2019; Sheahan et al.

2020).[15]

Identification of the average effect of NREGS (α_1) requires that, conditional on the full set of controls, changes in the outcome post-treatment must be due to the program, on average, and not to another omitted variable. But (α_1) is not the main quantity of interest here; I also note that for the FD specification it is identified using just one period.

1.5.4 Effect of NREGS on risk index of crop choice

First, I verify that the three measures of risk in district crop mix have skill in predicting aggregate crop yields. If these measures are correlated with actual aggregate risk, higher values of RICC should lead to yield losses after a bad rainfall shock while increasing yields after a good rainfall shock. I test this idea in equation 1.5. I expect $\theta_1 > 0$ since higher risk with normal weather years should be correlated with higher returns, $\theta_2 < 0$ and $\theta_3 < 0$ since higher risk with bad rainfall shocks or higher than normal temperatures should reduce yields, and $\theta_4 > 0$ since higher risk with good rainfall should increase yields.

[15] A literature on the bias of TWFE has developed recently, including that arising from differential timing. Callaway and Sant'Anna (2021) provide a framework to eliminate some of the bias arising from differential timing. Their approach relies on parallel trends conditional on baseline covariates, similar to this setting. But it is unable to test for increased weather sensitivity of aggregate crop yields after NREGS since there are no never-treated units in this setting. Therefore, I cannot conduct the whole analysis using their approach as it would limit the analysis to years until 2007, the year before the last phase of the program was implemented (since the CS estimator only makes use of untreated units as counterfactuals). Another complication in this setting is that treatment effect might evolve over time. Nevertheless, I plan to utilize their R *did* package to explore whether conditional parallel trends are likely to hold through a pre-trend check with data until 2007.

$$
\begin{aligned}
log(RVY_{dy}) = \theta_1\, log(RICC)_{dy} + \theta_2\, log(RICC)_{dy} * HDD_{dy} \\
+ \theta_3\, log(RICC)_{dy} * Low\_Rain_{dy} + \theta_4\, log(RICC)_{dy} * High\_Rain_{dy} \\
+ D_d + Y_y + \epsilon_{dy}
\end{aligned}
\tag{1.5}
$$

Next, I test whether increased weather sensitivity of crop yields after NREGS can be explained by increased agricultural risk embedded in the district crop mix, using the three measures of Risk Index of Crop Choice (RICC) separately. Since this paper is interested in understanding crop choice as a driver of yield volatility, I focus on the RICC-SD that is constructed using the second moment of the pre-program distribution. The RICC-mean and RICC-SD measures are strongly correlated, reflecting the fact that higher revenue crops also have higher volatility. Changes in RICC come from changes in area weights across crops. Equation 1.6 below presents the regression specification.

$$
\begin{aligned}
log(RICC_{dy}) = \alpha_2\, NREGS_{dy} + \tilde{\gamma_2}\, Weather_{dy} + \tilde{\beta_2}\, NREGS_{dy} * Weather_{dy} + \\
\lambda_d^p * t + \phi_2^1\, Z_d * t + \phi_2^2\, Weather_{dy} * Z_d + D_d + Y_y + \epsilon_{dy}
\end{aligned}
\tag{1.6}
$$

The literature on farmer investment decisions shows that they pay close attention to signals of what the weather is likely to be, including weather forecasts, before making investment decisions (Rosenzweig and Udry 2014). The Indian subcontinent receives most of its rainfall in the monsoon season that runs from June-September. One of the most important signals that farmers look at is early season rainfall;

this is the period in which sowing/planting of most major crops takes place, and weather in the rest of the season affects crop growth but not crop choice. However, rainfall and temperatures in the monsoon season affect soil moisture for the Rabi (winter) season crops. Therefore, I control for planting season (June-July) weather in contrast to the whole monsoon season used in the yield regressions.

Since crop choice is baked in before full weather realization, the main coefficient of interest is α_2, the average effect of NREGS on aggregate risk in crop choice with a normal *planting* season weather. The coefficients $\tilde{\beta}_2$ and $\tilde{\gamma}_2$ are informative of any changes in crop choice that occur as a result of planting season weather that is a signal for the full weather realization; these coefficients are to be interpreted differently from $\tilde{\beta}_2$ and $\tilde{\gamma}_2$. As with the regressions in the previous section, I successively introduce trends that vary by phase and initial district characteristics, and interact weather with initial characteristics. I do not conduct a first difference analysis as for Revenue Value of Yield; the FD specification may not be as informative about α_2 since the first difference of the NREGS dummy turns on only once when the program starts.

1.5.5 Effect of NREGS on agricultural wages

In order to shed light on the labor market channel that could cause aggregate yields to become more sensitive to NREGS, I run the following regression.

$$
\begin{aligned}
log(wage)_{idy} = \alpha_3 * NREGS_{dy} + \tilde{\gamma_3}\, Weather_{dy} + \tilde{\beta_3} NREGS * Weather_{dy} \\
+ \quad \eta H_{idy} + \lambda_d^p * t + \phi_3^1\, Z_d * t + \phi_3^2\, Weather_{dy} * Z_d \\
+ M_m + D_d + Y_y + \epsilon_{idy}
\end{aligned}
\tag{1.7}
$$

The coefficient α_3 captures the average effect of NREGS on agricultural wages, while β_3 captures changes in the wage sensitivity to weather shocks post-NREGS. Since the early NREGS districts were selected partially based on district characteristics that could be correlated with the individual-level outcome, I utilize a similar strategy to the regressions for the revenue value and risk index by flexibly controlling for trends in baseline values and weather as well as allowing NREGS phase-wise trends. Identification of α_3 and β_3 requires similar assumptions to that for the Revenue Value of Yield regressions.

Since these are individual-level regressions I also include a month-of-year dummy that controls for any seasonal variation in wages. The vector H_{idy} contain the usual controls for gender, age group, education levels, caste, religion and marital status that are included in a Mincer-type regression.

Imbert and Papp (2015) utilized data from 2004 and 2007 to conduct a standard difference-in-difference analysis of the effect of NREGS on wages for casual labor. I extend their analysis by estimating this effect using employment and wage data from 2003, 2004, 2005, 2007, 2009 and 2011.

1.6 Results

1.6.1 Initial results

I start with the impact of weather shocks on crop yields in Table 1.2. I confirm results found in the literature showing that HDD and rainfall shocks are important determinants of agricultural productivity. In particular, an extra heating degree day in the growing season reduces aggregate revenue value of yield by 1.9%, a low rainfall shock reduces yield by 7.5% and a high rainfall shock increases yields by 4.3%.

Next, I discuss the effect of weather shocks on measures of NREGS program activity, as proxied by three different variables. These results are shown in table 1.3. We see that a low rainfall shock increases the number of per capita person-days by 0.303 SD, the per capita number of households that work more than 100 days by 0.321 SD and the per capita expenditure on labor by 0.482 SD, although the first result is not statistically significant.[16] On the other hand, a high rainfall shock reduces these measures by 0.2 SD, 0.014 SD and 0.110 SD respectively (the second measure is not statistically significant).

However, the effect of HDD shocks is not to increase program activity, but rather to *decrease* activity by 0.211 SD, 0.042 SD (insignificant) and 0.093 SD (only significant at 10% level). The rainfall results confirm that negative (positive) agri-

[16] Results are robust to an IHS transform of the provision measures (to allow for zeros) rather than a standardization.

cultural productivity shocks lower (increase) average earnings and therefore increase (decrease) demand for NREGS. The HDD results suggest that bureaucrats pay more attention to proxies of agricultural productivity rather than knowledge of true productivity, since rainfall shocks may be easier to measure and understand than temperature deviations.

1.6.2 Main results

Table 1.4 presents the results for increased weather sensitivity of aggregate crop yields after NREGS. The main coefficients of interest are for the interaction of NREGS and weather variables. This table shows that a low rainfall shock after NREGS reduces yield further by 8.1% in the most demanding specification in column 4, and 10.4% in the first difference in column 5. Reassuringly, this result is consistent across all specifications. In contrast, high rainfall shocks do not change the sensitivity of yield, while NREGS also does not change the effect of heating degree days, even though the first difference specification suggests a decrease in the sensitivity to HDD.

Appendix table 1.A1 conducts an indirect test of the parallel trends assumption by using a placebo treatment. I move the NREGS indicator up by 5 years, as if the program had first started in 2001 and not 2006. I estimate equation 1.3 on this data from 1998-2008. The coefficient on the *Low Rain X NREGS* variable is statistically indistinguishable from zero, providing reassuring evidence that the results from Table 1.4 are attributable to NREGS and not to omitted variables.

Table 1.5 provides further evidence that the increased rainfall sensitivity is due to NREGS by adding an interaction term between weather, NREGS indicator and the lagged provision of NREGS to results in Table 1.4. I include all three measures of NREGS provision separately in different regressions using the most demanding specification of column 4 in Table 1.4. While I interpret the coefficients on *Low Rain X NREGS* as being causal, there may be need for caution in interpreting the coefficient on the lagged interaction term *Low Rain X NREGS Provision* in a causal manner. From Table 1.3, a negative rainfall shock increases provision on average; if this shock also affects agricultural productivity in the next year this correlation would be picked up in the interaction term. However, Kaur (2019) does not find a dependence of agricultural productivity on lagged rainfall shocks in the Indian context. I plan to test this with my data as well.

Going back to table 1.5, the estimate of the impact of a low rainfall shock post-NREGS on crop yield, conditional on average level of NREGS provision within district in the previous year, is between -5.3% and -9.9% in columns 1-6, and mostly measured precisely. The coefficients on *Low Rain X NREGS Provision* (*High Rain X NREGS Provision*) are estimates of the additional low (high) rainfall sensitivity from a 1 SD increase in NREGS provision in the previous year. The additional low rainfall sensitivity from 1 SD higher lagged provision ranges from -1.8 % to -8.4 %, measured quite precisely.[17] Additional high rainfall sensitivity is small and insignificant for two columns but precisely measured from 0.023% to

[17] Results are robust to an IHS transform of the provision measures (to allow for zeros) rather than a standardization.

0.03% in four other columns.

These results indicate that the decision to provide NREGS after a negative rainfall shock has repercussions into the next year. First, there is a trade-off between its consumption smoothing benefits after a negative rainfall shock and the negative effect this provision may have on agricultural yields if there is another negative shock in the next year. On the other hand, conditional on a positive rainfall shock in the next year, there is a positive complementarity between between higher provision in a given year and next year's crop yields.

1.6.3 Mechanisms

I now turn to the potential mechanism of risk in crop choice that has been hypothesized to explain increased weather sensitivity. Figure 1.2 shows that crops with higher mean revenue in a region are also likely to be subject to higher revenue volatility. Table 1.6 presents the results of a formal test, based on 1.5, of whether the constructed Risk Indices of Crop Choice capture real-work agricultural risk in the form of crop revenue volatility. The coefficients on *RICC* in column 1 and 2 for both RICC-Mean and RICC-SD show that higher risk is strongly correlated with higher returns in normal rainfall years, with elasticities of 1.17 and 1.073 respectively.

On the other hand, coefficient on *RICC* for column 3 suggest RICC-Skew does not predict returns during average years. Coefficients on the interaction term *RICC X*

HDD are negative as expected for all three measures and strongly significant for RICC-Mean and RICC-SD, with small elasticities of -0.003. Tellingly, given the importance of rainfall shocks in the Indian agricultural context, the coefficients on *RICC X Low Rain* show a consistent pattern of reductions in crop revenue, with elasticities for RICC-Mean and RICC-SD of -0.005. The coefficient on RICC-Skew should not be interpreted as an elasticity since the IHS transform behaves like a log transform only with values above 10 or so for the raw variable (Bellemare and Wichman 2020).[18] But the coefficient is in the same direction as the other two. Similarly, the coefficients on *RICC X High Rain* are positive and significant for all three measures of aggregate risk. Now that we have seen that these measures of aggregate risk have skill in predicting crop revenue, we should be able to tease out whether NREGS shifts district crop mix toward higher risk, higher revenue crops that are also positively skewed. Table 1.7 tests these hypotheses.

The three columns of table 1.7 are estimated for each of the three measures of risk in crop choice using the most demanding specification from equation 1.6 including phase-wise time trends, baseline controls interacted with time trends and weather shocks interacted with baseline controls. The coefficient on $NREGS$ is of interest and estimates the average impact on each measure of risk in crop choice during normal planting season years. From columns 1 and 3, there do not seem to be any changes in the average district crop mix toward higher mean yield crops (col 1), or more positively skewed crops (col 3). The coefficient in column 2 is significant at

[18] The max skewness in the data is 3.33

the 10% level and suggests a shift in average district crop mix toward more volatile crops. This result suggests that NREGS may have increased average risk in crop choice during normal planting seasons by 0.08%, as measured by the RICC-SD. Such a result could explain part of the increased sensitivity of yields to a negative rainfall realization.

However, there are a couple of reasons this result may not explain the increased sensitivity. The elasticity of Revenue Value of Yield to RICC-SD in Table 1.6 column 2, row 3 is -0.007%. Therefore, an increase in risk of 0.08% can only explain a (0.08*0.007) = 0.0056% reduction in yield. The actual reduction in yield with a low rainfall shock after NREGS is abut 10%.

Secondly, in table 1.A2, I test whether the results in table 1.7 satisfy the indirect parallel trends assumption. Column 2 of table 1.A2 shows that the placebo program dummy increases RICC-SD by 1.3%. Since the program was not in place at this time, the parallel trends assumption seems to be violated in this case, and a pre-existing trend may be driving the result observed in table 1.7. So, the true effect of NREGS on the Risk Index of Crop Choice may be even smaller, and it is unlikely that this particular mechanism is the cause of the increased rainfall sensitivity of crop yields after NREGS.

Finally, I provide some evidence in favor of the labor market channel. The NREGS literature documents clear increases in wages for unskilled labor, mainly due to general equilibrium effects that reallocate workers away from the private market (Muralidharan et al. 2016; Sukhtankar 2017). But beyond this level effect on unskilled wages, the coefficient on *Low Rain X NREGS* in table 1.A3 demonstrates that NREGS also reduces wage sensitivity to low rainfall shocks (Santangelo 2019). This mechanism can increase labor costs for farmers who are already dealing with a negative productivity shocks, thereby exacerbating the net productivity losses.

1.7 Conclusion

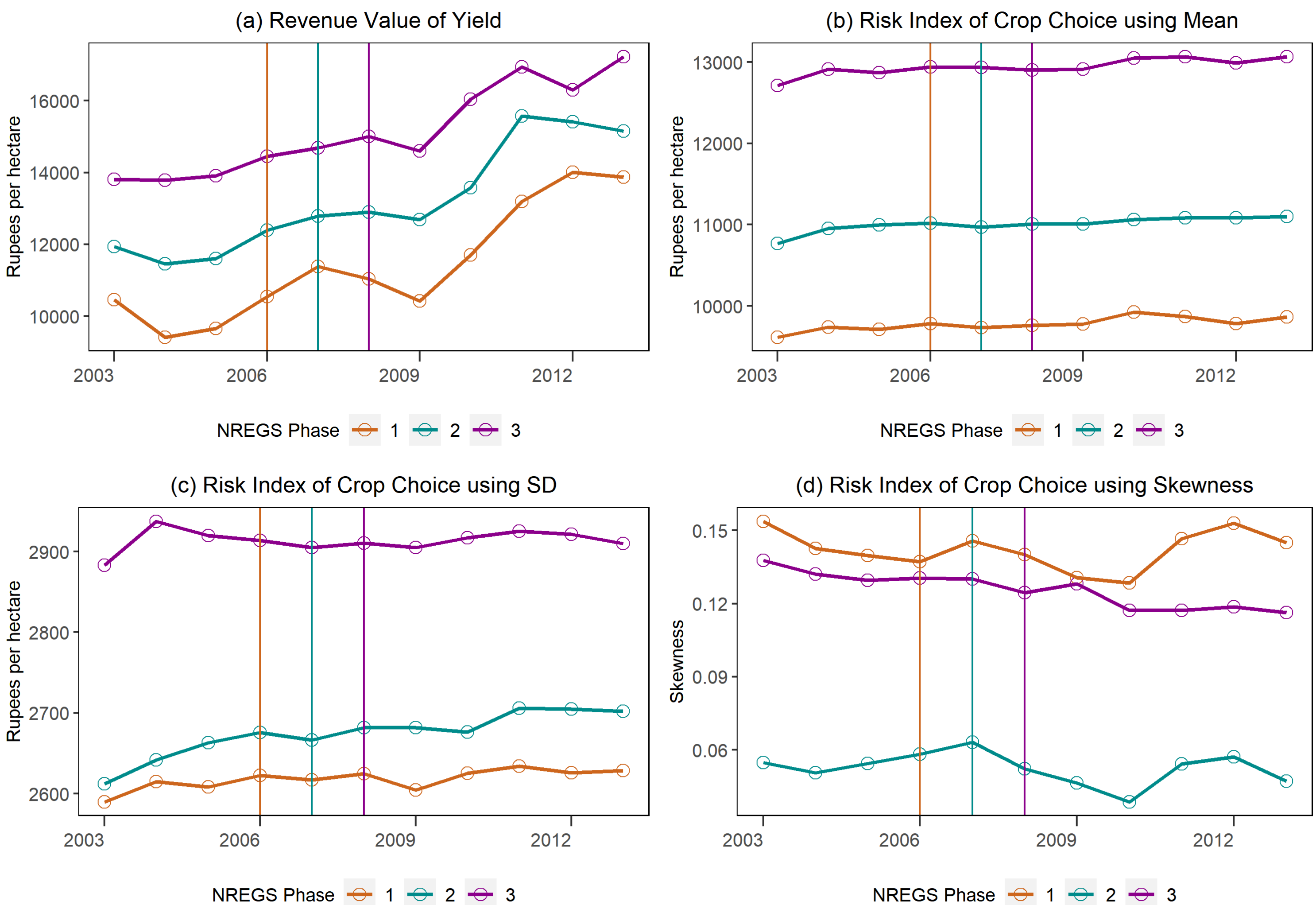
The stated purpose of NREGS was to improve livelihood security of the rural poor by providing them with income from manual work on demand. The program succeeded in its main goal of improving earnings for the poor and helped reduce poverty (Sukhtankar 2017). But I document that the program has economically large implications for the volatility of agricultural output by making aggregate yields more sensitive to negative rainfall shocks. I construct novel measures of aggregate risk in district crop mix to analyze whether the increased sensitivity is due to higher risk-taking by some farmers in response to the social insurance properties of NREGS. I show that these measures are meaningful because they have skill in predicting the volatility of yields. Using these measures of risk, I argue that the increased crop sensitivity cannot be explained by higher aggregate risk in the district crop mix. It is important to note that these measures of risk in crop choice may not capture other agricultural risk such as increased use of costly inputs such as fertilizers or machinery that could also reduce yields during a bad year if farmers are also credit constrained at the same time. Further research is necessary to understand those mechanisms better.

I also provide evidence consistent with the literature on the labor market channel that NREGS makes agricultural wages less elastic to rainfall shocks. This inhibits any pro-cyclical correction that would reduce labor costs during harvest and thereby prevent additional crop losses. However, the welfare implications of this are not straightforward; small and medium farmers who are net sellers of labor on the agricultural market indirectly benefit through higher earnings from the wage effect as well as directly through provision of NREGS, transferring agricultural risk to larger farmers. In a utilitarian sense, this might be a net welfare gain since the marginal utility of consumption of smaller farm-households is higher than that of larger ones, and they are also more numerous. There also exist positive (negative) complementatarities at the intensive margin of higher provision of NREGS to deal with rainfall shock in a given year, and subsequent aggregate yield if a positive (negative) rainfall shock is realized next year.

The aggregate food security implications of workfare programs, especially in the context of climate change, are not very well-understood. This paper sheds some light on this question for India, pointing to the labor market channel as the most important channel. By transferring yield risk from smaller to larger farmers while simultaneously increasing incomes of the latter, NREGS improves consumption smoothing for the poorest. But larger weather shocks in the future might lead to much larger yield losses and aggregate food security concerns, especially if the ability to store or source essential food grains and other staples is low. The literature on the implications of climate change for agriculture discusses trade across regions with less-correlated changes in climate as a potential solution (Costinot et al. 2016). However, shocks are likely more correlated within-country. Another avenue would be structural transformation such that fewer people depend on agriculture for livelihoods, and only the most productive farmers stay in agriculture (Suri 2011). If these farmers are able to consolidate land, they may be even more productive, given that larger farmers are less credit constrained and more able to invest in technologies (Foster and Rosenzweig 2017). But, land market consolidation is difficult given land market frictions, and other barriers to structural transformation generally, which could also make increased rainfall sensitivity of yields from NREGS more salient.

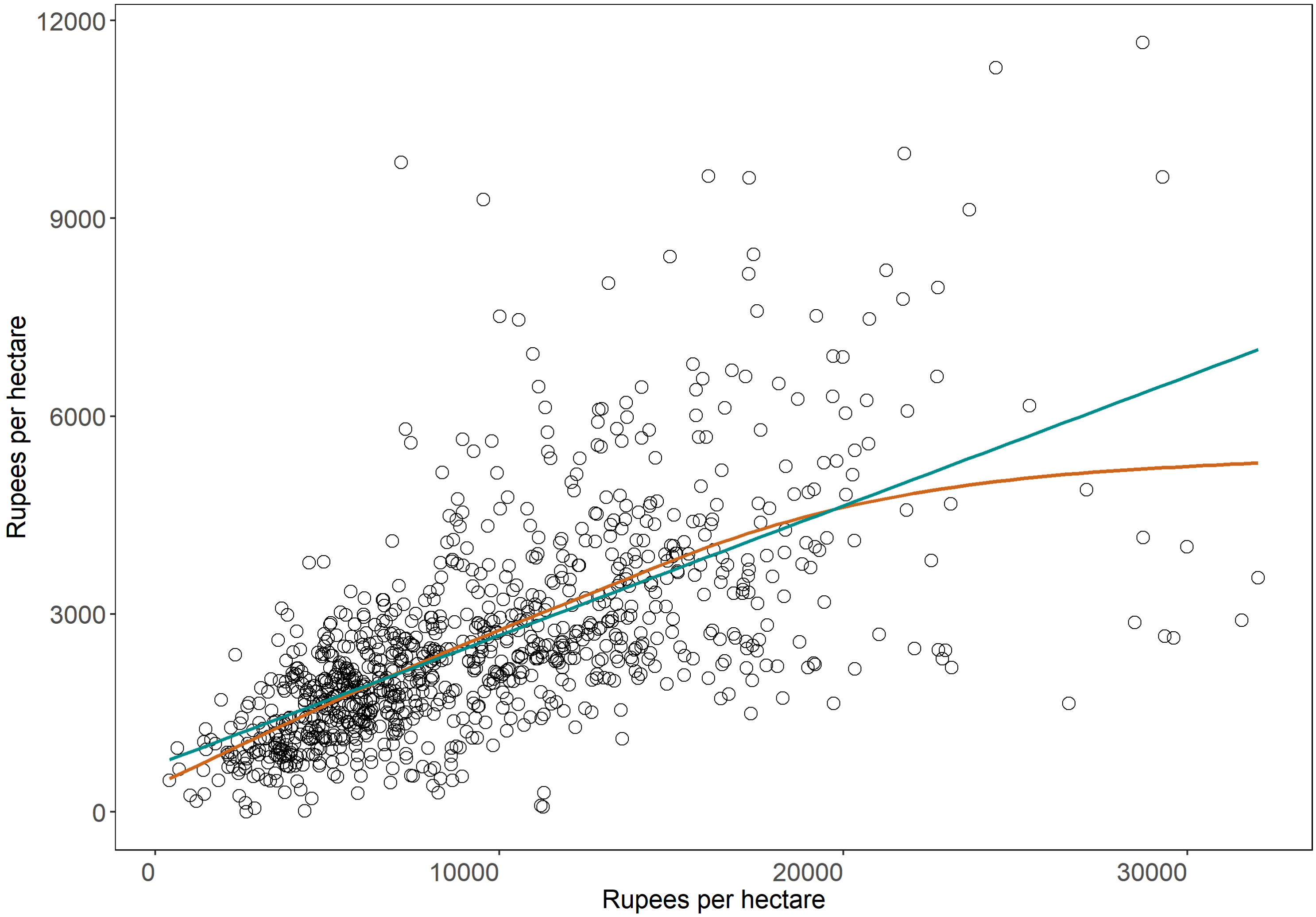
1.8 Figures and Tables

Figure 1.1: Trends in Revenue Value of Yield and Risk Indices of Crop Choice ↩



Notes: The first three moments for the distribution of revenue value of yield are separately calculated for each of the 16 crops in each of the 96 regions using data on yearly revenue value of yield between 1990-2002. The Risk Index of Crop Choice for each moment is the calculated by yearly crop area-weighted average of the specified moment of this distribution for the region within which the district falls (5.5 districts within each region on average). Panels (b), (c) and (d) display the yearly average of these risk indices, separately for each NREGS phase. Vertical lines provide NREGS start year by phase.

Figure 1.2: Mean vs SD of pre-2003 revenue value of yield distribution ↩



Notes: Each dot represents a crop-region. The Mean (y-axis) and SD (x-axis) for this distribution are calculated for each of the 16 crops in each of the 96 regions separately using data on yearly revenue value of yield between 1990-2002. Linear and cubic fits are also shown within the figure.

Table 1.1: Summary Statistics

	Early - 2006			Mid - 2007			Late - 2008		
Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD
Panel A: Pre-program variables used to determine district NREGS phase									
Share of lower castes	169	0.35	0.17	98	0.25	0.1	200	0.2	0.1
Ag product per capita (Rs/person)	169	5563	5930	98	7053	8731	200	12196	13910
Casual daily labor wage (Rs)	169	26.77	17.81	98	26.2	18.65	200	21.86	18.21
Panel D: NREGS Variables									
NREGA dummy	1859	0.73	0.45	1078	0.64	0.48	2210	0.55	0.5
HH worked > 100 days (Count/person)	1715	0.04	0.08	1015	0.02	0.05	2109	0.01	0.05
Person-days worked (Count/person)	1715	2.11	4.64	1015	1.74	3.95	2109	1.35	3.47
Labor Expenditure per capita (Rs/person)	1715	0.61	0.67	1015	0.42	0.54	2109	0.29	0.51
Panel B: Agricultural Outcomes									
Revenue value of yield (Rs/ha)	1859	11427	4427	1078	13226	4980	2210	15159	6306
RICC-Mean (Rs/ha)	1859	9778	3229	1078	11005	3503	2210	12941	4703
RICC-SD (Rs/ha)	1859	2618	790	1078	2674	815	2210	2913	1238
RICC-Skew (Skewness)	1859	-2.66	1.3	1078	-2.63	1.46	2210	-2.7	1.97
Panel C: Weather Variables - monsoon season (Jun-Oct)									

continued

Table 1.1: Summary Statistics (Continued)

Demeaned HDD (Degree-days)	1859	0.33	0.59	1078	0.31	0.5	2189	0.45	0.91
Low Rain Dummy	1859	0.23	0.42	1078	0.25	0.43	2210	0.22	0.42
High Rain Dummy	1859	0.27	0.44	1078	0.27	0.44	2210	0.27	0.44
Panel D: Weather Variables - planting season (Jun-Jul)									
Demeaned HDD (Degree-days)	1859	-9.82	11.95	1078	-11.34	8.99	2189	-10.93	8.34
Low Rain Dummy	1859	0.27	0.45	1078	0.27	0.44	2210	0.27	0.44
High Rain Dummy	1859	0.26	0.44	1078	0.24	0.43	2210	0.25	0.43
Panel E: Wage outcomes									
Ag daily labor wage (Rs)	53762	84.67	57.18	34689	94.69	78.94	53362	111.61	90.26

Notes: Time-invariant pre-program variables in Panel A are calculated from the National Sample Survey 2004 and Population Census of 2001. RICC in panel B refers to the Risk Index of Crop Choice. Yearly NREGS provision variables in Panel B are from 2009-2015. Weather variables in Panels C and D are calculated for from the Google Earth Engine. Individual wage outcomes in Panel E are the NSS. ↩

Table 1.2: Impact of weather shocks on aggregate and individual crop yields

	Dependent variable: log(RVY)					
	(1)	(2)	(3)	(4)	(5)	(6)
HDD	−0.019**	−0.016*	−0.027***	−0.052***	−0.132	−0.035*
	(0.008)	(0.010)	(0.006)	(0.019)	(0.100)	(0.019)
Low Rain	−0.075***	−0.073***	−0.039***	−0.023	−0.031	−0.128***
	(0.016)	(0.023)	(0.013)	(0.016)	(0.036)	(0.025)
High Rain	0.043***	0.055***	0.043***	0.026	0.035	−0.011
	(0.012)	(0.017)	(0.012)	(0.019)	(0.029)	(0.018)
Crop	Aggregate	Rice	Wheat	Sugarcane	Cotton	Groundnut
Observations	10,134	9,421	8,353	4,136	4,136	6,322
R2	0.787	0.780	0.829	0.545	0.544	0.570
District and Year FE	X	X	X	X	X	X

Notes: Estimation on data from 1990-2013. RVY refers to Revenue Value of Yield. Each column presents estimates for either the aggregate revenue value of yield for all crops weighted by area planted, or individual crop revenue value of yield. HDD refers to heating degree days above 25C. Low and High Rain are dummies for rainfall below and above 20th or 80th percentile of historical rainfall. Weather variables are for the **monsoon period (June-October)**. Conley standard errors using a cutoff of 1000 km and arbitrary autocorrelation up to 5 years are reported. All columns include district and year fixed effects. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 1.3: NREGA provision in response to weather shocks

	Dependent variable: Standardized per capita provision		
	(1)	(2)	(3)
HDD	-0.211^{**}	-0.042	-0.093^{*}
	(0.083)	(0.064)	(0.050)
Low Rain	0.303	0.321^{***}	0.482^{***}
	(0.203)	(0.111)	(0.117)
High Rain	-0.205^{*}	-0.014	-0.110^{*}
	(0.113)	(0.078)	(0.060)
NREGS Provision var	Person days	Num HH > 100 Days	Labour Expenditure
Observations	2,117	2,117	2,117
R2	0.708	0.680	0.787
District and Year FE	X	X	X

Notes: Estimation on data from 2009-2013. Each column presents estimates for a standardized measure of per capita NREGS provison. HDD refers to heating degree days above 25C. Low and High Rain are dummies for rainfall below and above 20th or 80th percentile of historical rainfall. Weather variables are for the **monsoon period (June-October)**. All provision variables are converted to per capita and then standardized. Conley standard errors using a cutoff of 1000 km and arbitrary autocorrelation up to 5 years are reported. All columns include district and year fixed effects. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 1.4: Impact of NREGA on Weather Sensitivity of Aggregate Yields

	Dependent variable: log(RVY)				
	(1)	(2)	(3)	(4)	(5)
NREGA	−0.003	0.013	0.018	0.016	0.042
	(0.028)	(0.030)	(0.030)	(0.029)	(0.026)
HDD X NREGA	0.023	0.027	0.021	0.023	0.043^{**}
	(0.017)	(0.017)	(0.016)	(0.016)	(0.021)
Low Rain X NREGA	-0.081^{**}	-0.081^{**}	-0.085^{**}	-0.081^{**}	-0.104^{***}
	(0.040)	(0.040)	(0.039)	(0.037)	(0.035)
High Rain X NREGA	0.007	0.012	0.005	0.006	−0.020
	(0.027)	(0.027)	(0.027)	(0.028)	(0.032)
Observations	5,132	5,132	5,126	5,126	4,665
R2	0.810	0.811	0.813	0.815	0.086
Trend X Phase		X	X	X	
Trend X Controls			X	X	
Weather X Controls				X	
First Difference					X
District and Year FE	X	X	X	X	X

Notes: Years 2003-2013. RVY refers to aggregate Revenue Value of Yield for all crops weighted by area planted. HDD refers to heating degree days above 25C. Low and High Rain are dummies for rainfall below and above 20th or 80th percentile of historical rainfall. Weather variables are for the **monsoon period (June-October)**. Conley standard errors using a cutoff of 1000 km and arbitrary autocorrelation up to 5 years are reported. All columns include district and year fixed effects. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 1.5: Impact of Provision on Weather Sensitivity of Aggregate Yields

	Dependent variable: log(RVY)					
	(1)	(2)	(3)	(4)	(5)	(6)
NREGA	0.019	0.009	0.012	0.058**	0.045*	0.057**
	(0.029)	(0.029)	(0.030)	(0.026)	(0.025)	(0.027)
Provision	−0.019*	0.011	−0.018	0.010	−0.021	0.023
	(0.011)	(0.011)	(0.013)	(0.009)	(0.019)	(0.020)
HDD X NREGA	0.018	0.024	0.022	0.012	0.021	0.018
	(0.015)	(0.016)	(0.016)	(0.018)	(0.018)	(0.018)
Low Rain X NREGA	−0.072*	−0.080**	−0.053	−0.085**	−0.099***	−0.070*
	(0.037)	(0.037)	(0.040)	(0.039)	(0.038)	(0.041)
High Rain X NREGA	−0.005	0.007	−0.009	−0.047	−0.034	−0.052*
	(0.029)	(0.029)	(0.029)	(0.031)	(0.031)	(0.031)
HDD X Provision	0.013*	0.006	0.005	0.009**	0.012	0.004
	(0.007)	(0.005)	(0.005)	(0.004)	(0.008)	(0.008)
Low Rain X Provision	−0.018*	−0.056***	−0.040**	−0.021**	−0.084***	−0.043**
	(0.009)	(0.022)	(0.020)	(0.010)	(0.030)	(0.019)
High Rain X Provision	0.030***	−0.004	0.023**	0.028***	0.006	0.027**
	(0.009)	(0.010)	(0.012)	(0.009)	(0.010)	(0.012)
Provision Variable	HH100Days	PDays	LabExp	HH100Days	PDays	LabExp
Observations	4,872	4,872	4,872	3,945	3,945	3,945
R2	0.820	0.818	0.820	0.112	0.107	0.111
First Differences				X	X	X
District and Year FE	X	X	X	X	X	X

Notes: Years 2003-2013. RVY refers to aggregate Revenue Value of Yield for all crops weighted by area planted. Provision is the lagged value of the per capita standardized NREGA measure listed under each column. HDD refers to heating degree days above 25C. Low and High Rain are dummies for rainfall below and above 20th or 80th percentile of historical rainfall. Weather variables are for the **monsoon period (June-October)**. Conley standard errors using a cutoff of 1000 km and arbitrary autocorrelation up to 5 years are reported. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 1.6: Does Risk Index of Crop Choice predict Crop Yields?

	Dependent variable: log(RVY)		
	(1)	(2)	(3)
RICC	1.170^{***}	1.073^{***}	-0.109
	(0.160)	(0.149)	(0.108)
RICC X HDD	-0.003^{**}	-0.003^{**}	-0.025
	(0.001)	(0.001)	(0.021)
RICC X Low Rain	-0.005^{*}	-0.007^{**}	-0.120^{***}
	(0.003)	(0.003)	(0.046)
RICC X High Rain	0.005^{***}	0.006^{***}	0.075^{**}
	(0.002)	(0.002)	(0.035)
RICC Var	log(RICC-Mean)	log(RICC-SD)	ihs(RICC-Skew)
Observations	5,132	5,132	5,132
R2	0.825	0.818	0.806
District and Year FE	X	X	X

Notes: Years 2003-2013. RVY refers to aggregate Revenue Value of Yield for all crops weighted by area planted. RICC refers to the Risk Index of Crop Choice. Each column presents an estimate using a different RICC that is constructed by crop area-weighting one of the first three moments of the the pre-2003 crop revenue distribution. HDD refers to heating degree days above 25C. Low and High Rain are dummies for rainfall below and above 20th or 80th percentile of historical rainfall. Weather variables are for the **monsoon season (June-October)**. Conley standard errors using a cutoff of 1000 km and arbitrary autocorrelation up to 5 years are reported. All columns include district and year fixed effects. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 1.7: Impact of NREGA on Risk Index of Crop Choice

	Dependent variable:		
	log(RICC-Mean)	log(RICC-SD)	ihs(RICC-Skew)
	(1)	(2)	(3)
NREGA	0.001	0.008*	0.002
	(0.005)	(0.004)	(0.006)
HDD X NREGA	0.001*	0.001**	0.0002
	(0.0003)	(0.0002)	(0.0002)
Low Rain X NREGA	0.0003	−0.006	−0.010*
	(0.006)	(0.004)	(0.006)
High Rain X NREGA	0.009	0.001	−0.005
	(0.008)	(0.006)	(0.006)
Observations	5,126	5,126	5,126
R2	0.982	0.985	0.982
District and Year FE	X	X	X

Notes: Years 2003-2013. RICC refers to the Risk Index of Crop Choice. Each column presents an estimate using a different RICC that is constructed by crop area-weighting one of the first three moments of the the pre-2003 crop revenue distribution. HDD refers to heating degree days above 25C. Low and High Rain are dummies for rainfall below and above 20th or 80th percentile of historical rainfall. Weather variables are for the **planting period (June-July)**. Conley standard errors using a cutoff of 1000 km and arbitrary autocorrelation up to 5 years are reported. All columns include district and year fixed effects. *p<0.1; **p<0.05; ***p<0.01. ↩

1.9 Appendix

Table 1.A1: Placebo Impact of NREGA on Weather Sensitivity of Agg. Yields

	Dependent variable: log(RVY)				
	(1)	(2)	(3)	(4)	(5)
NREGA	0.020	0.019	0.019	0.020	−0.028
	(0.027)	(0.026)	(0.026)	(0.024)	(0.031)
HDD X NREGA	0.014	0.013	0.011	0.013	0.003
	(0.021)	(0.022)	(0.021)	(0.020)	(0.032)
Low Rain X NREGA	−0.014	−0.014	−0.013	−0.008	0.033
	(0.039)	(0.039)	(0.039)	(0.037)	(0.035)
High Rain X NREGA	−0.043	−0.043	−0.041	−0.033	−0.029
	(0.041)	(0.042)	(0.043)	(0.043)	(0.042)
Observations	5,092	5,092	5,091	5,091	4,625
R2	0.816	0.816	0.816	0.818	0.157
Trend X Phase		X	X	X	
Trend X Controls			X	X	
Weather X Controls				X	
First Difference					X
District and Year FE	X	X	X	X	X

Notes: Years 1998-2008. RVY refers to aggregate Revenue Value of Yield for all crops weighted by area planted. HDD refers to heating degree days above 25C. Low and High Rain are dummies for rainfall below and above 20th or 80th percentile of historical rainfall. Weather variables are for the **monsoon period (June-October)**. Conley standard errors using a cutoff of 1000 km and arbitrary autocorrelation up to 5 years are reported. All columns include district and year fixed effects. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 1.A2: Placebo Impact of NREGA on Risk Index

	Dependent variable:		
	log(RICC-Mean)	log(RICC-SD)	ihs(RICC-Skew)
	(1)	(2)	(3)
NREGA	0.011	0.013*	−0.006
	(0.007)	(0.007)	(0.006)
HDD X NREGA	0.0005*	0.0002	−0.0004
	(0.0002)	(0.0002)	(0.0003)
Low Rain X NREGA	−0.006	−0.009	0.001
	(0.006)	(0.006)	(0.005)
High Rain X NREGA	0.002	−0.00002	0.013**
	(0.007)	(0.006)	(0.005)
Observations	5,091	5,091	5,091
R2	0.983	0.984	0.981
District and Year FE	X	X	X

Notes: Years 1998-2008. RICC refers to the Risk Index of Crop Choice. Each column presents an estimate using a different RICC that is constructed by crop area-weighting one of the first three moments of the the pre-2003 crop revenue distribution. HDD refers to heating degree days above 25C. Low and High Rain are dummies for rainfall below and above 20th or 80th percentile of historical rainfall. Weather variables are for the **planting period (June-July)**. Conley standard errors using a cutoff of 1000 km and arbitrary autocorrelation up to 5 years are reported. All columns include district and year fixed effects. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 1.A3: Impact of NREGA on wages for hired agricultural labor

	Dependent variable: log(Wage)			
	(1)	(2)	(3)	(4)
NREGA	0.017	0.008	0.012	0.012
	(0.027)	(0.027)	(0.027)	(0.027)
HDD X NREGA	−0.004	0.006	0.0002	−0.002
	(0.016)	(0.015)	(0.016)	(0.016)
Low Rain X NREGA	−0.056*	−0.055*	−0.053*	−0.057**
	(0.029)	(0.028)	(0.028)	(0.028)
High Rain X NREGA	−0.043	−0.037	−0.043	−0.044
	(0.028)	(0.028)	(0.027)	(0.027)
Observations	129,845	129,845	129,845	129,845
R2	0.387	0.387	0.388	0.389
Trend X Phase		X	X	X
Trend X Controls			X	X
Weather X Controls				X
District, Month and Year FE	X	X	X	X

Notes: Years 2003, 2004, 2005, 2007, 2009 and 2011. Outcome variable is log of individual daily wage earnt while working on manual labor tasks in the private market. HDD refers to heating degree days above 25C. Low and High Rain are dummies for rainfall below and above 20th or 80th percentile of historical rainfall. Weather variables are for the **monsoon period (June-October)**. Standard errors are clustered at the district level. All columns include district, year and month fixed effects. *p<0.1; **p<0.05; ***p<0.01. ↩

Chapter 2

The Long-Term Impact of Air Pollution on Aggregate Productivity and Spatial Inequality

2.1 Introduction

Developing countries face an unprecedented environmental challenge in the form of air pollution today. The worst affected country is India, where average particulate matter (PM2.5) levels at 7 times the prescribed World Health Organization standards may be reducing life expectancy by 6 years (Greenstone 2021). Could these high levels of air pollution at a relatively low level of development affect aggregate growth patterns? This paper considers the implications of high air pollution in some Indian cities for aggregate productivity and spatial inequality.

The literature documents substantial negative labor productivity impacts of contemporaneous air pollution (Graff Zivin and Neidell 2012; Chang et al. 2019; Fu et al. 2021). High levels of air pollution in some cities could lead to a long-term reallocation of economic activity away from these cities because they are less productive and liveable, leading to aggregate losses from spatial misallocation of labor. If these cities were poorer to begin with, such a reallocation would lead to further divergence in per capita incomes within the country.

Local air pollution can be affected by interjurisdictional externalities such as wildfires that consistently increase pollution in some cities (Burke et al. 2021). I focus on the use of unregulated agricultural fires as the source of air pollution in downwind Indian cities. Low levels of economic development may lead to such polluting activity, especially in the context of weak regulatory capacity (Jayachandran 2022; Besley and Persson 2009). By documenting the impact of distant sources of local air pollution on long run aggregate outcomes, this paper underlines the importance of regulating such sources for aggregate growth patterns.

Wildfires are known to affect air pollution hundreds of kilometers away (Rogers et al. 2020). I begin the analysis by documenting that agricultural fires in India similarly increase PM2.5 concentrations in distant, downwind cities. To quantify the long-term aggregate productivity impacts of fires-driven air pollution, it is necessary to account for any spatial reallocation of economic activity in response to changes in pollution and productivity across cities. Therefore, I employ a Quantitative Spatial Equilibrium (QSE) framework of location choice (Redding and Rossi-Hansberg 2017) in the spirit of Rosen (1974) and Roback (1982), incorporating air pollution externalities across geographical units into the model.

The first mechanism that could cause spatial misallocation is the direct impact of air pollution on labor productivity. Higher fire exposure leads to higher air pollution and lower labor productivity, thereby leading to lower incomes and fewer workers being attracted to the location. Cities that are closer to and in the path of prevailing winds from sources of fires could be more productive if these fires were reduced. The income elasticity of migration governs how strongly location choice responds to income differences across potential locations - a lower elasticity implies that higher incomes need to be paid on average to induce marginal workers to migrate. The extent to which reducing fires increases aggregate output depends on this income elasticity.

Secondly, if workers respond by migrating away from more polluted cities due to lower amenity value, that could reduce agglomeration forces and make those cities less productive. I model this mechanism through a pollution elasticity of migration that determines how strongly worker location choice responds to the level of pollution itself. Previous work has shown that workers in rich countries do care about air quality in their location choice (Banzhaf and Walsh 2008), while recent evidence suggests that workers in China also respond in similar ways (Khanna et al. 2021).

The model equilibrium conditions lead to a labor supply equation across locations that can be used to estimate both the income and pollution elasticities of migration together, controlling for origin fixed effects.

[1] The pollution elasticity of migration is not significantly different from zero in the OLS while the income elasticity is precisely estimated and is 3 times larger than the magnitude of the pollution elasticity.

I also estimate a parameter that determines the prevalence of fires in agriculture, representing the sum total of agricultural practices and government policies. I utilize a district-level panel data set of rice area under cultivation and fire count at the annual level to quantify this elasticity. I calibrate parameters that control local pollution generation from non-agricultural production (Fu et al. 2021) and labor productivity losses from pollution (Adhvaryu et al. 2022).

With these parameters in hand, I conduct model policy counterfactuals to reduce the prevalence of agricultural fires in the states of Punjab and Haryana that are

[1] To allay any concerns about residual factors that could be correlated with real income or pollution, I also instrument for PM2.5 concentrations with exposure to upwind fires, and for own-district income with distance-weighted neighbor income following Tombe and Zhu (2019). Due to concerns about the strength of the second instrument, I do not present the IV results as the main estimates currently, although further work is underway. I also plan to use city-level in-migration data and variation in PM2.5 driven by new thermal power plants constructed recently. This will enaable the utilization of another measure of migration and a different source of identification, apart from also considering education-specific effects of pollution on migration decisions.

by far the main source of these fires. The benefits of this policy are to increase total GDP for the country by approximately 1.22%, and total welfare by 1.29%. This policy can be implemented through paying farmers not to burn, subsidizing machinery that avoids setting fields on fire, or simply through better enforcement on current regulations that outlaw any agricultural fires, each one of them with different associated costs. Most of the GDP gains come from improving labor productivity by reducing air pollution in North India from the fires. Sorting allows marginal workers in the South and East to move toward the newly more productive and liveable North, and therefore reallocate GDP. The additional gain in total GDP from better allocation of labor is small, perhaps due to the locations not trading with each other in the model.

While net GDP and welfare increase, the Gini coefficient of GDP reduces by 0.23% while the Gini for welfare increases by 0.27%. This implies a reduction in spatial inequality of GDP but an increase in welfare inequality. Intuitively, this result comes about because air pollution from fires predominantly affect the relatively poorer Northern states of India. The reallocation of GDP toward the North due to the policy increases its relative GDP vis-à-vis the rest of the country, thus reducing GDP inequality. But the higher population in the North also increases congestion effects (such as competition for housing), while reducing congestion in the rest of the country. Thus, while welfare increases in each location relative to baseline, it increases more in the rest of the country, explaining the increase in welfare inequality.

This paper contributes to the recent literature modeling environmental outcomes within QSE models (Balboni 2021; Khanna et al. 2021; Heblich et al. 2021), by incorporating spatial air pollution externalities over geographical units. There is a wider literature using general equilibrium Rosen-Roback models to measure the effect of various policies such as state taxes, internal trade frictions, housing constraints, infrastructure and land misallocation on aggregate growth.[2]

[2]See Fajgelbaum et al. (2019) for state taxes, Redding and Turner (2015) for internal trade frictions, Hsieh and

Secondly, while more stringent regulation (Burgess et al. 2019), use of technology (Assunção et al. 2022) and higher resource allocation to monitoring and enforcement (Duflo et al. 2018) can improve environmental outcomes in developing countries, this paper's key insight is that facilitating Coasian bargaining with suitable compensation to farmers in Punjab and Haryana could lead to productivity gains in other states. This regulatory failure may be down to poor state capacity (Jayachandran 2022) or due to inappropriate regulation (Lipscomb and Mobarak 2017; Kahn et al. 2015). Finally, I provide some evidence of a zero or small migration response to pollution in India, relative to estimates for China (Freeman et al. 2017; S. Chen et al. 2022; Khanna et al. 2021), although this estimate is imprecise.

The rest of the paper is structured as follows. Section 2 describes the data used in the paper while section 3 presents the context and motivates the QSE model by estimating the extent of pollution externalities from agricultural fires. Section 4 presents the model of general equilibrium while section 5 describes the estimation of the parameters governing equilibrium. Section 6 describes the results from model counterfactuals, and section 7 concludes.

2.2 Data and Measurement

2.2.1 Air quality

An important consideration for air quality data is complete geographical coverage. Whereas ground-level monitoring station coverage in India is extremely sparse (Greenstone and Hanna 2014), satellite imagery-based products provide complete coverage. Secondly, ground-level observations may be susceptible to manipulation (Greenstone et al. 2022; Ghanem and Zhang 2014). Therefore, the main source of data on air quality is Hammer et al. (2020), a gridded reanalysis product of global surface PM2.5 concentrations at a resolution of 0.01° that should be much

Moretti (2019) for housing constraints, Duranton et al. (2015) for land misallocation, Ahlfeldt et al. (2015) for infrastructure less susceptible to such manipulation. This product combines satellite imagery data on Aerosol Optical Depth with state-of-the-art chemical transport models, and calibrates the output to global ground-based observations. This product has been used in the litrature as the main measure of PM2.5 in settings where ground level observations are sparse such as China (Khanna et al. 2021). It is easy to aggregate the gridded product to the necessary resolution for analysis at pixel, city or district level. Forthcoming sections will detail the aggregation procedure for each analysis.

2.2.2 Agricultural fires

The burning of residue from crop harvest is referred to as agricultural fires. There are no representative ground-level observations of this phenomenon, but the National Aeronautics and Space Administration (NASA) agency of the United States produces the Fire Information for Resource Management System (FIRMS) product that is widely used to identify such fires. This product provides information on daily, pixel-level fire detection data across the world. FIRMS provides a few related products: a Near-Real Time (NRT) fires using the MODIS instrument aboard Terra and Aqua satellites, a quality-controlled standard product from the same instrument but with a 2-3 month lag and another NRT product using the VIIRS instrument from the Suomi-NPP and NOAA-20 satellites. The main difference between the first two and the third is the resolution of the data. MODIS products are at 1 km resolution and are available from 2000 (with higher reliability from 2002 onward when the Aqua satellite was launched) whereas VIIRS products are at 375 m but only available from 2012.

Since I require data from before 2012, I am unable to use the higher-resolution VIIRS-based product. The primary analysis here utilizes the MODIS quality-controlled standard product which differs from the NRT data in that corrections are made to the imprecise location of the Aqua satellite in the NRT data. Imagery data from Aqua and Terra satellites is available at least four times daily for each pixel on Earth and is processed using a NASA algorithm to isolate a ground-level fire signal from other signals such as solar flares.[3]

I combine this data with information on land use from the European Space Agency Climate Change Initative's land cover map (version 2.07).[4] This allows the subset of fires that is found on agricultural land to be separated from natural forest fires since this paper is interested in agricultural fires. I aggregate and resample the land cover data which is at a resolution of 300 m to the fire data grid (at 1 km resolution), with an indicator for agricultural land use as the main output from this process. All fires are then masked based on this indicator variable to find the subset of agricultural fires.

2.2.3 Migration

The source of data on migration in this paper is the Population Census of India 2011.[5] The estimating equation that characterizes spatial equilibrium in the quantitative model requires migration shares between all pairs of locations. Therefore, we need data on such flows to estimate pollution and income elasticities of migration that determine spatial equilibrium. The population census provides tabulations of the number of people canvassed in each district by district of birth. Following Bryan and Morten (2019), I construct migration shares from these tabulations using district of birth as the origin and district of enumeration as the destination. This measure of migration shares should capture all migrants except for those who were not present in their destination at the time of enumeration for idiosyncratic reasons.

The census provides two separate tabulations of locations of birth and location of enumeration (the person's location at the time of interview for the census). Firstly, table D-01 provides data on the number of people enumerated in a given district who were born in other districts of the same state. This allows construction of

[3] Further information on these products is available at https://firms.modaps.eosdis.nasa.gov

[4] Data is available at https://cds.climate.copernicus.eu/cdsapp/#!/dataset/satellite-land-cover

[5] See https://censusindia.gov.in/census.website/data/census-tables

within-state migration shares across districts. Secondly table D-11 provides data on the number of people who were enumerated in a given district but were born in a different state of the Union of India. In order to construct the complete data set on migration shares across *all* districts of India, I need to allocate data from table D-11 on the total number of workers who were born in a given state but moved to a district outside that state, to the various districts in the state of origin. This data is not publicly available.

In order to do this, I utilize information on the out-migration tendency of districts from table D-01 by calculating each district's share of out-migrants within the same state. Then I assume that these within-state out-migration shares are the same for out-migration to districts in other states. This allows me to allocate the number of out-migrants to a district outside the state, to each district of the state of origin. To the extent that certain districts send out more migrants, whether within state or outside, this method would capture migration shares from a district in a given state to another district outside that state, with some error.

2.2.4 GDP and price data

The quantitative model is estimated on state level data for which GDP within agriculture and non-agriculture is required. Such data are published by the Planning Commission of India for 2004-2011. Price data is necessary to construct real GDP. I utilize the National Sample Survey household consumption module for 2011-12 to construct individual item price data, and construct weights for these items following standard methods used to construct price indices from the same data source.

2.2.5 Meteorological data

Hourly wind data are used to construct exposure to agricultural fires for every origin-destination pixel pair. Details of the methodology follow in section 3 below. The source of these wind data is the European Center for Medium Range Weather

Forecasting (ECMWF) ERA5 family of global gridded reanalysis datasets.[6] Reanalysis data combine ground-level observations and satellite data with Chemical Transport Models that represent physical and chemical processes in the atmosphere to produce reliable and complete coverage for the world. Since ground-level observations are particularly sparse in developing countries, these reanalysis data are widely used in the literature on climate and air pollution in Economics (Auffhammer et al. 2013). Hourly wind speed and direction data are taken from the ERA5-Land hourly dataset which is available at a resolution of 0.1°. These are combined with daily agricultural fires at the pixel level to construct the fire exposure variable, as described in section 3 below. Finally, I also construct temporal averages for weather variables including rainfall, temperature and relative humidity from this data set to be used as controls in the regression analysis.

Secondly, to construct data on thermal inversions, I utilize the ERA5 data set on hourly temperature at a resolution of 0.25° and at various pressure levels through the atmosphere.[7] Atmospheric pressure reduces with altitude as the amount of air exerting pressure decreases. Temperature usually also decreases with altitude above sea level, but occasionally an upper layer of the atmosphere is at a higher temperature than a lower level, hindering the diffusion of air pollutants through the atmosphere. This phenomenon is referred to as a thermal inversion and is widely used as an instrument for air pollution in the literature (Dechezleprêtre et al. 2019; Bondy et al. 2020). Data on temperatures at various pressure levels is used to identify when a thermal inversion by matching the land surface altitude with the correct pressure level. Since the land surface of India has varied altitude, I utilize a digital elevation model to guides the correct pressure level at the surface and the construct thermal inversion based on temperature at the pressure layer above.

[6] Data is available at https://cds.climate.copernicus.eu

[7] See https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels

2.3 Context and Motivation

2.3.1 Causes of agricultural fires

The primary use of fires in Indian agriculture today is to clear the field of leftover residue from harvesting a crop, before sowing and planting the next crop (Shyamsundar et al. 2019); this differs from slash-and-burn agriculture that is practiced in parts of Africa and Indonesia (Andini et al. 2018). Figure 2.1 provides a map of where agricultural fires are most prevalent in India; fires are concentrated in the North-Western states of Punjab and Haryana ('PH').

The wheat growing states of Punjab and Haryana also grow rice, thereby creating a rice-wheat system. In these rice-wheat systems, rice is cultivated during the monsoon or "Kharif" season (June-November) while the wheat crop is cultivated in the winter or "Rabi" season (December-April). The rice crop harvesting process leaves a residue in the field that makes planting of wheat in early Rabi season difficult; fires are a cheap way to remove this residue. Wheat must also be planted in the first weeks of winter in order to get optimal yields. The short duration between the rice harvest in late October and the optimal wheat planting window in early November also incentivizes farmers to burn the residue so that they can prepare the fields for wheat planting faster.

The rice-wheat system has its roots in the Green Revolution of the 1960s. Until then, North-Western India was a primarily wheat-growing region with little rice consumption or production locally. The advent of the Green Revolution brought with it many institutional innovations from the Indian State that increased agricultural productivity substantially across India. In the states of Punjab and Haryana, this took the form of massive subsidies for tubewells which could be used to access shallow groundwater to irrigate fields that did not have access to the pre-existing large canals systems built by the colonial British empire. This newfound access to groundwater allowed farmers to diversify their crop portfolio during the monsoon months by allowing the cultivation of water-intensive rice crop. The state of Punjab contributed less than 1% of India's rice in 1961; by the late 1990s this figure was up to 10%, even as total rice output across India also increased substantially. The use of fires to clear rice residue started in the 1990s. The earliest observations of fires from the NASA FIRMS database starting in 2002 clearly demonstrate that North-western India already had a disproportionate share of fires in Indian agriculture.

2.3.2 Quantifying air quality externalities from agricultural fires

The stock of pollution in a location is a result of both local and external fires. Air pollution modelers in India describe emissions from agricultural fires affecting air quality up to a thousand kilometers away. On the other hand, air pollution from only certain non-agricultural sources such as thermal power plants is transported over such long distances.[8] Therefore, I capture the contribution of agricultural emissions E^a_o from a 1x1 degree origin pixel o on air quality in destination pixel d, at a distance of $dist_{od}$,[9] in calendar year y in equation 2.1

$$\omega_{ody} = \left(\sum_{t=Jan-1}^{Dec-31} \frac{windfrac_{odt} * E^a_{ot}}{dist_{od}} \right) \tag{2.1}$$

$windfrac_{odt}$ is the daily average fraction of time that the wind at o blows towards d on day t. In order to calculate $windfrac_{odt}$, I start by assigning each hourly wind observation in o on day t into one of 36 bins of 10 degree span each, based on the wind direction that hour (true north is 0 degree as in the figure). I then construct the wind speed-weighted fraction of time the wind was blowing in each of these 36 bins by aggregating hourly observations for day t. $windfrac_{odt}$ is then calculated by summing up wind fraction for the bins which are in the direction of d from o as

[8] See "Basics 004 - diffused vs. point and local vs. non-local sources" at https://urbanemissions.info/blog-pieces/whats-polluting-delhis-air.

[9] The distance elasticity is assumed to be -1 but can also be estimated using non-linear least squares

shown in figure 2.2.

Daily "fire exposure" is then calculated by multiplying daily wind fraction $windfrac_{odt}$ by daily emissions E^a_{ot} at origin and dividing by $dist_{od}$. Yearly fire exposure from pixel o to d is then the sum of daily fire exposures. With these measure of yearly fire exposure for each origin-destination in hand, I then construct total fire exposure Ω_d for destination d as the sum of exposures ω_{od} from all origins o.

$$\Omega_d = (\sum_{o=1}^{N} \omega_{od}) \tag{2.2}$$

The matrix Ω allows me to construct an Econometric Transport Model (ETM) for agricultural fires by summarizing the exposure of any given location to agricultural fires in *all* other locations. It flexibly accounts for daily changes in the count of fires or wind patterns at the origin, the most important factors in determining the exposure of downwind districts. Figure 2.4 demonstrates visually that this fire exposure metric positively correlates with PM2.5 across Indian districts.

In order to estimate the causal effect of yearly local fire count E^a_{dy} and external fire exposure Ω_{dy} on average yearly PM2.5 concentration Z_{dy}, I run the following regression on 1x1 degree pixels using panel data from 2002-2018.

$$log(Z_{dy}) = \gamma\, log(E^a_{dy}) + \tau\, log(\Omega_{dy}) + Y_y + D_d + \epsilon_{1dy} \tag{2.3}$$

The identification assumption for γ is that local pollution levels do not affect local fires directly. The assumption for τ is that local pollution levels do not affect upwind fires. These are likely to be satisfied, given that most regulations on agricultural fires are not implemented.

Table 2.3 describes the results of this exercise. Since fires may be zero on any given pixel, I run three measures of fire count, including log of fire count, inverse hyperbolic sine (ihs) of fire count and log of (1+fire count). An increase in external fire exposure of 1% increases PM2.5 by 0.22-0.24%. Since fire exposure is a weighted sum of all external fires, increases in exposure can come from a small increase in the fire count in a nearby location that is upwind, or by a large increase in another upwind location that is much farther.

On the other hand, local fire count does not seem to significantly affect local PM2.5 once fire exposure is accounted for, although the estimates on fire count are always positive. I attribute this to agricultural fires being predominantly employed in Punjab and Haryana (North-Western India) whereas PM2.5 levels are high all over India, and especially in the North. This result underlines that the spillover effects of external agricultural fires are more important to local pollution than local fires themselves. In order to quantify how important fire exposure is, I note that the within-R2 increases from 0.011 to 0.116, suggesting that around 10% of the yearly variation in PM2.5 within districts can be explained by fire exposure.

2.4 Spatial Equilibrium Model with Pollution Externalities

This section incorporates the econometric pollution transport model into a canonical quantitative models of economic geography (Redding and Sturm 2008; Ahlfeldt et al. 2015) to investigate the effects of reducing agricultural fires on aggregate productivity. The reader is referred to Redding and Rossi-Hansberg (2017) for a survey of this literature.

2.4.1 Worker preferences

There are $\bar{L}_o$ workers in location o to begin with. Workers have preferences over agricultural (C_d^a) and non-agricultural (C_d^n) goods. Preferences are assumed to take a Cobb-Douglas form.

$$u_{od} = \epsilon_{od}(C_d^a)^{\alpha}(C_d^n)^{1-\alpha} Z_d^{\lambda} B_d M_{od}$$

where C_d^a, C_d^n are consumption of final agricultural and non-agricultural goods. B_d are fixed amenities that can include climate and other institutional features. Housing supply is assumed to be inelastic and therefore absorbed in amenities B_d. ϵ_d is an idiosyncratic preference shifter that captures preferences for location d. ϵ_d is i.i.d across workers, locations and sectors and drawn from a Frechet distribution given by the CDF $F(\epsilon) = e^{-\epsilon^{-\eta}}$. A larger value of ϵ_{od} implies that worker is particularly attached to location d and would not move even with large wage differentials or high pollution levels. This would capture real world features such as strong local ties, for example.

Z_d is the level of pollution in location d. If workers have preferences over clean air, locations will be characterized by compensating differentials for pollution, with elasticity given by λ. If $\lambda < 0$ then pollution lowers utility whereas if $\lambda = 0$ then pollution does not directly affect worker utility. In the empirical section, I show that $\lambda = 0$ and therefore air quality does not directly enter utility. Therefore, I will omit this term from utility from now onwards.

A given worker in location o decides whether to move to destination location d; there are N locations to choose from. Workers can move freely across sectors within a location, but movement across locations is costly. Migration cost M_{od} from origin o to destination d represents physical costs of migration and also any salient differences in culture etc. This is motivated by the fact that about 80% of migration in India is within the state, an entity that shares a common language and cultural features.

Labor income in location d is given by wage w_d and is the same for both sectors as there is free movement of labor given the free mobility across sectors assumption. Workers choose the location where they receive highest utility, subject to moving costs. If the indirect utility function for the worker is represented by V_{od}, then the worker chooses d over d' if $V_{od} > V_{od'}$. Indirect utility function is given by

$$V_{od} = \epsilon_d B_d M_{od} \left(\frac{w_d}{(P_d^a)^\alpha (P_d^n)^{(1-\alpha)}} \right)$$

Let I_d be the real income such that

$$I_d = \frac{w_d}{(P_d^a)^\alpha (P_d^n)^{(1-\alpha)}}$$

This formulation allows us to write the migration share from o to d, π_{od} as follows, where we have made use of the properties of the Frechet distribution

$$\pi_{od} = \frac{l_{od}}{\bar{L}_o} = \frac{[I_d B_d M_{od}]^\eta}{\sum_{k=1}^{N} [I_k B_k M_{ok}]^\eta} \tag{2.4}$$

Equilibrium number of workers in d is given by $l_d = \sum_o l_{od}$. Local demand D_d^j for goods in sector $j \in \{a, n\}$ is pinned down by Cobb-Douglas shares.

$$D_d^a = \alpha w_d l_d \qquad\qquad D_d^n = (1-\alpha) w_d l_d$$

2.4.2 Production and general equlibrlium

Each location d produces a homogeneous good y_d^j in sector j using a linear technology with labor l_d^j and TFP A_d^j. Each worker supplies one unit of inelastic labor. TFP varies across locations and may be affected by pollution Z_d and agglomeration forces. Labor is assumed to be perfectly mobile across sectors.

$$y_d^j = A_d^j l_d^j$$

Markets are perfectly competitive. Therefore, the price of each good equals marginal cost.

$$p_d^j = \frac{w_d}{A_d^j}$$

All goods are produced and consumed locally so there is no goods trade. Output within sector j is then determined purely by demand D_d^j. Labor shares in each sector are determined by the Cobb-Douglas utility. This comes from the requirement that share of income spent on good j must equal revenue from that good, i.e. $\alpha w_d l_d = p_d^a y_d^a$ and $(1-\alpha) w_d l_d = p_d^n y_d^n$.

$$l_d^a = \alpha l_d \qquad\qquad l_d^n = (1-\alpha) l_d$$

Local wages and agricultural prices are pinned down by setting $p_d^n = 1$. Total earnings are purely from labor income w_d and are spent locally. Equilibrium is determined by the vector $\{l_d\}$ and is reached by varying l_d^j across locations such that equations 2.5, 2.6 and 2.7 below are satisfied.

$$w_d = A_d^n \qquad\qquad p_d^a = \frac{A_d^n}{A_d^a} \tag{2.5}$$

$$\pi_{od} = \frac{l_{od}}{\bar{L}_o} = \frac{\left[\frac{w_d}{(p_d^a)^\alpha} B_d M_{od}\right]^\eta}{\sum_{k=1}^{N} \left[\frac{w_k}{(p_k^a)^\alpha} B_k M_{ok}\right]^\eta} \tag{2.6}$$

$$l_d = \sum_{o=1}^{N} \pi_{od} \bar{L}_o = l_d^a + l_d^n \tag{2.7}$$

Productivity A_d^a, A_d^n and pollution Z_d are endogenous in this model. Elasticities that depend on employment govern generation of pollution, productivity impacts of pollution and agglomeration effects. Equilibrium values of these endogenous variables therefore depend on employment and are a part of the solver described above. I provide more details on the endogenous productivity and pollution specifications later.

2.4.3 Productivity, agglomeration and pollution impacts

TFP can vary by sector due to fixed exogenous factors like soil quality, presence of rivers, or availability of raw materials like mineral ores; agglomeration forces and the effect of pollution on worker productivity. Equation 2.8 formalizes these ideas. $\bar{A}_d^j$ is exogenously determined productivity that does not respond to employment.

$$A_d^j = \bar{A}_d^j Z_d^{\beta^j} {l_d^j}^{\phi^j} \tag{2.8}$$

β^j determines how productivity responds to pollution; if $\beta^j < 0$, productivity is negatively affected by pollution. The stength of agglomeration forces that may arise from any potential non-excludable innovation (Arrow 1962) is captured by ϕ^j.

2.4.4 Endogenizing pollution from production

Pollution Z_d in location d is given by

$$Z_d = \bar{Z}_d (L_d^n)^{\psi} (E_d^a)^{\gamma} \Omega_d^{\tau} \tag{2.9}$$

The pollution elasticity of non-agricultural employment is given by ψ. This elasticity is taken from Fu et al. (2021) for China where it is estimated to be 1.36.[10] Ω_d is the contribution of external fire exposure to local pollution in location d.

$$\Omega_d = \left(\sum_{o=1}^{N} \frac{windfrac_{od} * E_o^a}{dist_{od}} \right) \tag{2.10}$$

[10] I will be estimating this elasticity for India

As discussed earlier, $windfrac_{od}$ is the fraction of time the wind blows from o to d; I consider the 10-year daily average wind patterns when constructing this quantity for the model. The distance between the centroids of o and d is $dist_{od}$. These are exogenous physical determinants of fire exposure and together determine how exposed district d is to emissions in origin o.

The measure of agricultural emissions in district d is E_d^a. As described in section 2.3, the main determinant of the extent of agricultural fires is the quantity of rice produced, mediated by institutional features of that particular state. The quantity of rice produced is also proportional to the area under rice cultivation. This motivates the following specification

$$E_d^a = (\theta_d L_d^a)^{\delta_d} \tag{2.11}$$

Here rice area is measured by the inclusion of a scaling term θ_d reflects the importance of rice in the district crop mix on agricultural employment. Thus $\theta_d L_d^a$ represents the potential residue that might be burnt. I operationalize θ_d by setting it to the percentage of average cultivated land area that is dedicated to rice. The location-specific elasticity δ_d captures the local institutional factors that motivate the actual decision to burn. Details on estimation of δ_d follow in the next section.

2.5 Estimation of Model Parameters

Table 2.1 shows the model parameters that are estimated in this paper. Apart from the productivity elasticity of pollution and the effect of non-agricultural output on pollution, I estimate all the other parameters of the model. This section will present the research design and discuss results for each of these parameters. Summary statistics for the data used in the various parameter estimation exercises are provided in table 2.2

2.5.1 Elasticity of emissions to labor in agriculture (δ)

Appendix figure 2.3 demonstrates that institutional differences drive the increased prevalence of agricultural fires in Punjab and Haryana rice-wheat system. Therefore, I allow δ_d to differs based on whether the district is in Punjab and Haryana, or in another state. Further, labor employed for rice cultivation is proportional to area cultivated under rice ($\theta_d L_d^a \propto RiceArea_d$), allowing me to estimate δ_d with a yearly panel of fires and area under rice cultivation.

I run the following regression on data from 2002-2011, where E_{dy}^a represents annual fire count and $RiceArea_{dy}$ is the total area under rice in district d and year y. District fixed effects D_d control for any fixed characteristics of the district that are correlated with agricultural fires, state-by-year fixed effects Y_{sy} allow us to account for any changes in state government policies (which is the relevant administrative level for agricultural policy in India) that affect agricultural fires. Standard errors are clustered at district and state-by-year to account for autocorrelation as well as agricultural policy changes that are made at the state level.

$$\begin{aligned} log(E_{dy}^a) = \delta_{ph}\, &log(RiceArea_{dy}) * \mathbb{1}(d \in PH) \\ &+ \delta_{nph}\, log(RiceArea_{dy}) * \mathbb{1}(d \notin PH) \\ &+ Y_{sy} + D_d + \epsilon_{2dy} \end{aligned} \tag{2.12}$$

The main assumption required to identify δ is that yearly changes in area under rice cultivation are not correlated with yearly changes in area under sugarcane crop, which is the only other crop that is associated with agricultural fires in India. This is likely to be the case as sugarcane is not a substitute for rice. Sugarcane also is not likely to compete with rice for land, given that rice is a staple crop that is the subject of policies such as minimum support prices from most state governments.

Table 2.4 confirms the result shown in the figure 2.3. The elasticity of fires to rice area is 5 times larger for districts in Punjab and Haryana, reflecting institutional features that were baked in decades ago and do not respond to higher pollution levels currently.

2.5.2 Elasticities of pollution from local fires and external fire exposure (γ, τ)

These two parameters have already been identified in section 3, where I estimated equation 2.3, which is a log-linear version of equation 2.9. The third parameter ψ that determines pollution from non-agricultural production is calibrated, as described in the calibration section.

2.5.3 Income and pollution elasticities (λ, η), and Migration costs (m_{od})

Given the equilibrium migration shares predicted by the quantitative model, the income and pollution elasticities can be estimated with data on migration shares across Indian districts that were described in the data section. Taking natural log of equation 2.4 gives us the following

$$
\begin{aligned}
log(\pi_{od}) = \eta\,(log(w_d)\; - \alpha\, log(P_d^a)) \qquad & [Real\ Income] \\
- log(\bar{V}_o) \qquad & [Origin\ option\ value] \\
+ \eta\lambda\, log(Z_d) \qquad & [Pollution\ amenity] \\
+ \eta\, log(B_d) \qquad & [Fixed\ amenity] \\
- \eta\, m_{od} \qquad & [Migration\ cost]
\end{aligned}
\tag{2.13}
$$

In order to estimate this equation on migration shares data, we need to specify migration cost M_{od}. I follow Khanna et al. (2021) in assuming that M_{od} take the form $M_{od} = exp(-m_{od})$, where m_{od} are parameterized such that migration costs are normalized and symmetric ($m_{oo} = 1$ and $m_{od} = m_{do}$).

$$m_{od} = \nu_1 log(dist_{od}) + \nu_2 \mathbb{1}(lang_{od})$$

Here $dist_{od}$ measures the physical distance between districts o and d while $\mathbb{1}(lang_{od})$ captures whether a different language is spoken in o and d.

$$\begin{aligned} log(\pi_{od}) = &\ \eta\,(log(w_d) - \alpha\, log(P_d^a)) \\ &+ \eta\lambda\, log(Z_d) \\ &- \eta\,\nu_1 log(dist_{od}) - \eta\,\nu_2\,\mathbb{1}(lang_{od}) \\ &+ F_o + \epsilon_{3od} \end{aligned} \tag{2.14}$$

where F_o is an origin fixed effect that absorbs $log(\bar{V_o})$ and the residual ($\epsilon_{3od} = \eta\, log(B_d) + \epsilon_{4od}$) contains differences in destination amenities and other idiosyncratic features that determine bilateral migration shares. I then follow the literature (Ahlfeldt et al. 2015; Bryan and Morten 2019; Khanna et al. 2021) in estimating amenities B_d from the residual through the use of destination fixed effects, $\widehat{\epsilon_{3od}} = F_d + \epsilon_{4od}$, such that $\widehat{F_d} = \eta\, log(B_d)$.

Table 2.5 presents the results from estimation of 2.14. The OLS results are presented in column 1. From a given origin district, a 1% higher PM2.5 level at a potential destination district is associated with 0.19% lower migration share to that district, whereas a 1% higher real wage at the destination district is associated with a 0.81% higher migration share to that district. However, the coefficient on *log(PM)* is not significantly different from zero. This may be the result of either noise in measurement of migration shares for outside-state districts, or reflect heterogeneous preferences due to characteristics such as income wherein only high income individuals move in response to pollution (Greenstone and Jack 2015).

On the other hand, coefficients on *log(PM)* and *log(wage)* may be biased if there are residual destination-specific factors such as housing costs that are correlated with PM2.5 or real wage, even after controlling for origin fixed effects and other covariates. The coefficient on real wage and PM2.5 are both biased toward zero since the true real wage and exposure to PM2.5 are both negatively correlated with housing costs, but the expected signs on the coefficients are positive and negative respectively.

The standard solution to this problem is to instrument for both *log(PM)* and *log(wage)*. I follow Tombe and Zhu (2019) in instrumenting for *log(wage)* by a distance-weighted average of the neighboring district real wages. I instrument for *log(PM)* using yearly fire exposure.[11] Columns 2 and 3 display the first stage results for each instrument separately whereas column 4 presents the 2SLS result. While the fire exposure instrument has a strong first stage Kleibergen-Paap F-stat of 190.89, the neighbor-wage instrument has an F-stat of only 10.541. But distance enters the fire exposure instrument directly; districts that are close to each other may have correlated residuals due to fixed factors that affect migration to certain districts such as a pleasant climate or pre-existing infrastructure. Also, when there are two endogenous regressors, the best practice in testing for weak instruments is not yet established (Andrews et al. 2019). For these reasons, I prefer the OLS estimates at present.

Finally, 1% higher distance to destination district reduces migration to that district by 1.78% while a district where people speak a different language have ~36% lower migration share, holding all else constant.

2.5.4 Calibration

A few parameters of the quantitative model are calibrated. The Cobb-Douglas share of expenditure on agriculture, α, is calculated from data to be 0.26. Next, the pollution elasticity of non-agricultural labor that governs emissions from other activity is taken from Fu et al. (2021), who estimate this elasticity for a panel of Chinese manufacturing firms in the 2000s. The estimate of 1.36 for this elasticity is the closest to the Indian context in the literature.

Third, the elasticity of labor productivity to PM2.5 in agriculture and non-

[11] I utilize the strength of each fire, measured using the Fire Radiative Power from the NASA FIRMS data base to maximize predictive power, given that a larger fire would emit more particulate matter.

agriculture (β^a, β^n) are crucial parameters. I take these from Adhvaryu et al. (2022) who estimate the productivity elasticity of PM2.5 on individual-level data that they collect for urban, indoor garment workers in the Indian state of Gujarat. Their estimate of -0.05 is lower than the -0.09 estimated by Chang et al. (2019) for indoor pear-packers at a factory in the US, but I prefer the estimate from India. Due to the lack of a better estimate for outdoor agricultural workers as well, I take -0.05 as a lower bound for those workers as well.

Lastly, the agglomeration elasticity ϕ^n of 0.076 is taken from Chauvin et al. (2016). They estimate the effect of population density on individual wages in Indian districts using National Sample Survey data, relying on historical population density that is a standard instrument in this literature.

2.5.5 Model quantities

Model quantities such as the fixed productivity component are inverted from data and the structure of the model. Details are provided in table 2.6.

2.6 Results

2.6.1 Main counterfactual

The main policy counterfactual involves reducing the elasticity of emissions in Punjab and Haryana (δ_{ph}) to equal the elasticity for the rest of the country, δ_{nph}. This seems like a reasonable government objective when considering fires in Punjab and Haryana, since there are indeed some fires in other states.[12] The exact nature of policies to achieve this equalization of elasticities would differ, including in their costs. For instance, a few ways to implement this policy would be through better enforcement of existing regulations, to pay farmers not to burn stubble, or

[12] A separate counterfactual might be to completely eliminate agricultural fires, for instance through technological and institutional innovations. This will be the subject of later inquiry.

through a subsidy to rent mechanical harvesters called Happy Seeders that obviate the need for fires to get rid of residue.[13] Associated costs with these policies are different, but the benefits are quantified in the figure below. None of these policies includes changing the district crop mix itself; therefore, these estimates can be seen as regulatory or institutional changes that drive air quality without the need for difficult transitions in agricultural crop mix.

Table 2.7 describes model counterfactual results from implementing the policy to reduce fires in Punjab and Haryana. I focus on two outcomes of interest: GDP and welfare. In this model without trade, welfare maps directly to real incomes. I discuss net changes in GDP and welfare, how sorting contributes to these changes, and end with results on how the policy affects spatial inequality of GDP and welfare.

Since I am unable to reject that the pollution elasticity is different from zero, I begin the discussion assuming that the pollution elasticity of migration $\lambda = 0$. Row 1 of table 2.7 shows that the policy increases net GDP by 1.22% and net welfare by 1.29%. In order to understand the mechanisms behind this result better, I present changes in GDP and welfare for each state, without and with sorting, in figure 2.5.

I focus on the counterfactual without sorting first. Panel (a) of figure 2.5 shows that while GDP increases in all states, the increase is larger closer to Punjab and Haryana (remembering that these two states lie at the Northwestern corner, where fires are largest as in figure 2.1). Panel (a) of figure 2.6 provides an insight into why this is the case. The policy reduces fires by more than 90% in these two states. This drastic reduction in fires reduces fire exposure relatively more in the North in panel (c) of figure 2.6, and therefore also reduces pollution relatively more in the North in panel (a) of figure 2.7. The main effect of reducing pollution without allowing people to move is to make North India relatively more productive, although all regions gain in productivity. These increases in productivity also decrease relative

[13] These machines harvest rice while sowing the wheat seeds into soil along with burring the stubble back into the soil.

prices, driving relatively higher increases in real incomes and welfare in North India.

Next, I consider the additional effects of allowing people to move on GDP and welfare. We have just noted that without allowing workers to move, the policy reduces pollution and increases productivity in the North. If workers are allowed to move, both these factors draw marginal workers toward the North from the South, East and Punjab. This reallocation of population (which is the same as labor in the model) is plotted in panel (c) of figure 2.7. This is accompanied by a reallocation of GDP from these areas toward the North, relative to the no-sorting counterfactual. However, this increased economic activity in the Northern states is counteracted by increased congestion, relative to the no-sorting counterfactual. These congestion effects reduce welfare gains for the North while increasing it for the rest of the country, relative to no-sorting. The net effect of the policy on GDP and welfare with sorting are plotted in panels (b) and (d) of figure 2.5. The additional change from the no-sorting counterfactual are plotted in panels (b) and (d) of figure 2.A1.

To reiterate, row 1 of table 2.7 shows that change in total GDP and net welfare from the policy, allowing for workers to move, is 1.22% and 1.29% respectively. Row 2 shows that sorting adds 0.39% to net welfare, accounting for 30% of the welfare gains. At the same time, sorting adds only 0.01% to GDP. The large additional increase in welfare from sorting may be explained by lower congestion forces in the South and East, pollution and housing being good examples. In a model with trade, where some workers could stay behind and enjoy the lower prices of goods produced in the North, these welfare gains from sorting may be lower; in the current model without trade, the only way workers can get lower prices is by migrating. The reason why sorting only adds 0.01% gains to total GDP could be that a model without trade at the state-level does not allow efficient allocation to locations that have comparative advantage in manufacturing.[14]

[14] I have new sectoral GDP data at the district level and will run model with trade at district level in a future iteration.

Finally, this discussion demonstrates that the net changes hide much heterogeneity. The third row of table 2.7 shows that the Gini coefficient of GDP reduces 0.23%, implying a reduction in spatial inequality of GDP. But the Gini of welfare increases by 0.27%; even though welfare increases everywhere, it increases relatively more in the South.

2.7 Conclusion

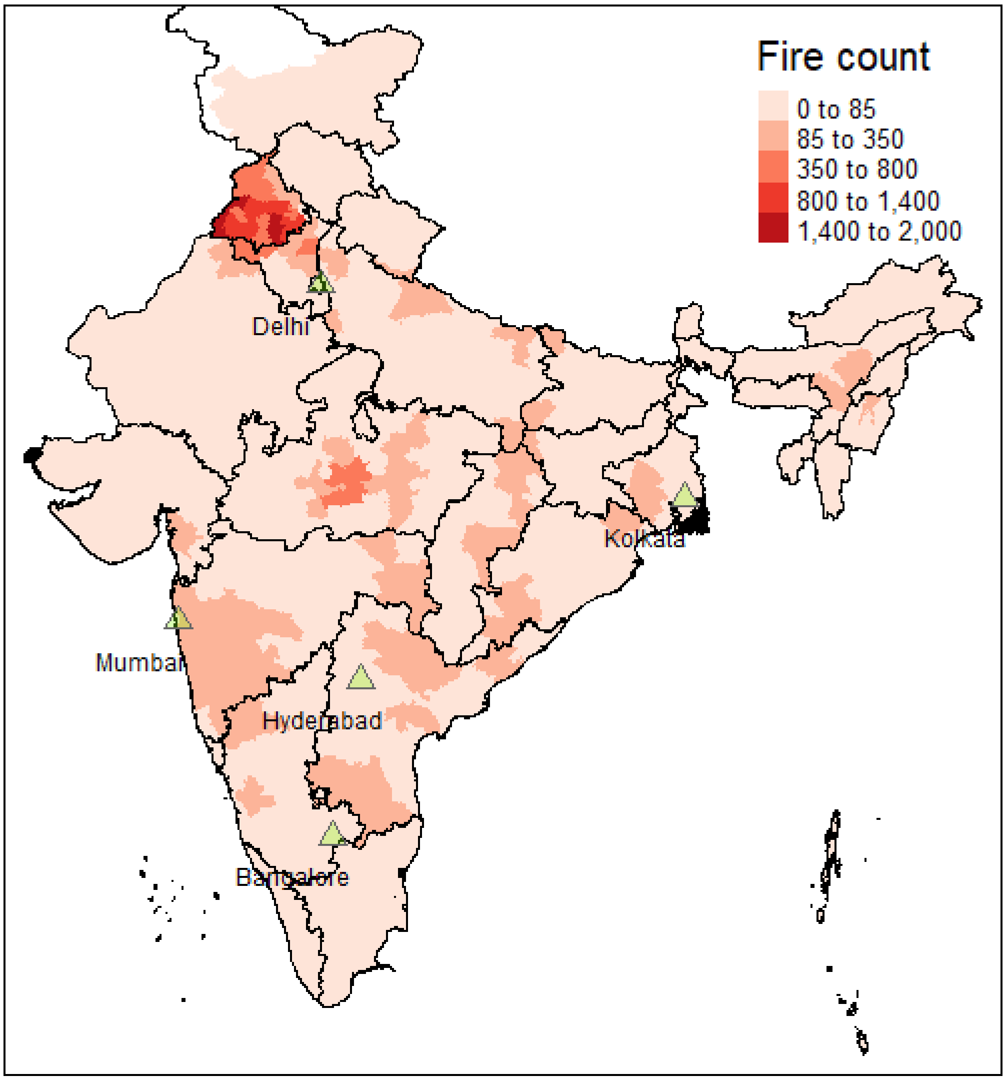
This paper studies the aggregate productivity effects of air pollution caused by agricultural fires in India. I develop an econometric transport model and demonstrate that agricultural fires originating outside district boundaries account for more than 10% of within-district annual variation in PM2.5, whereas local fires only account for ~1%. This pollution externality causes districts that could be more productive and employ more workers to be less productive. To quantify the gains from reducing these fires, I build a spatial equilibrium model of location choice, incorporating across locations. I estimate some of the parameters of this model. Firstly, I estimate that the pollution elasticity of migration, how much do workers respond directly to air pollution by moving away from it, is not significantly different from zero, although measured imprecisely. Secondly, I estimate the income elasticity of migration, the extent to which workers are responsive to relative changes in real income across locations, to be 0.81. Thirdly, I estimate a parameter governing the institutional determinants of agricultural fires in the model. This parameter differ between the states of Punjab and Haryana that account for more than 80% of all agricultural fires and in the rest of the country, representing different agricultural practices and policies.

I then conduct a policy counterfactual to change the institutional parameter to reflect a world in which residue burning practices in Punjab and Haryana look more like the rest of India, reducing the prevalence of fires by more than 90%. I find that this increases national GDP by ~1.22% and net welfare by 1.29%. The net gains in GDP are much higher than most estimates of the monetary cost of eliminating fires altogether in Punjab and Haryana. These gains in GDP are also higher for the poorer North India, since agricultural fires affected air pollution most in those states. On the other hand, while welfare increases in all states, larger increases are observed in the South and East from where population and GDP reallocated toward the North. The explanation for this result is that while reallocation of economic activity increases GDP and welfare, some of the welfare gains from reductions in pollution in North India are dissipated away by the congestion forces that come with increases in population. A note of caution in interpreting these results is that the quantitative model does not allow for trade, thereby negating one of the avenues by which a more efficient spatial allocation of labor could be achieved, with different implications for GDP and welfare changes.

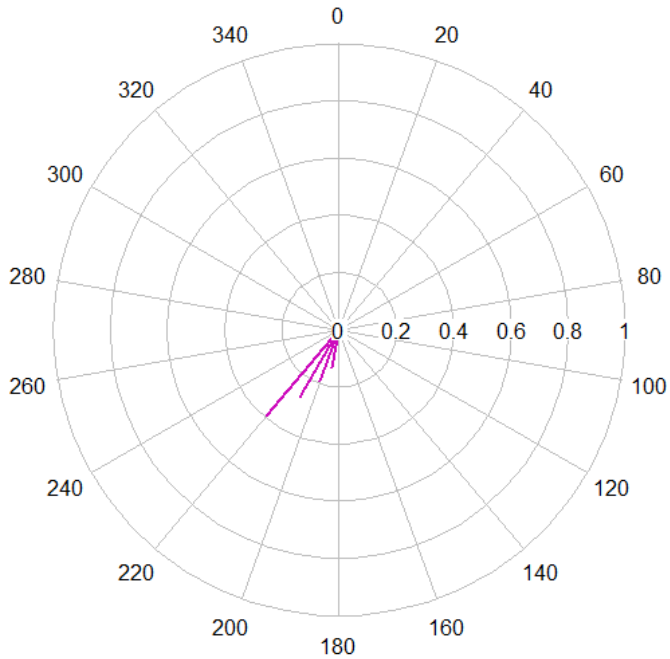
A significant finding of this paper is that workers in states that are far away from the major sources of agricultural fires and therefore not directly affected by pollution from those fires, indirectly suffer losses from these fires. If pollution from these fires were reduced, these workers would be able to take advantage of higher incomes and living standards in the North, not just existing workers in the North. This suggests the failure of a Coasian bargaining process wherein the rest of India could have compensated Punjab and Haryana in some way in return for a costly reduction in fires. This failure may be down to lack of regulation of the pollution externalities at the appropriate level (Banzhaf and Chupp 2012; Lipscomb and Mobarak 2017; Kahn et al. 2015). A deeper root cause could be that the regulatory capacity necessary to tackle such large environmental externalities do not exist in many developing countries (Jayachandran 2022).

2.8 Figures and Tables

Figure 2.1: Count of fires in Indian districts (2010) ↩



(a) Average wind directions at origin



(b) Direction from origin to destination

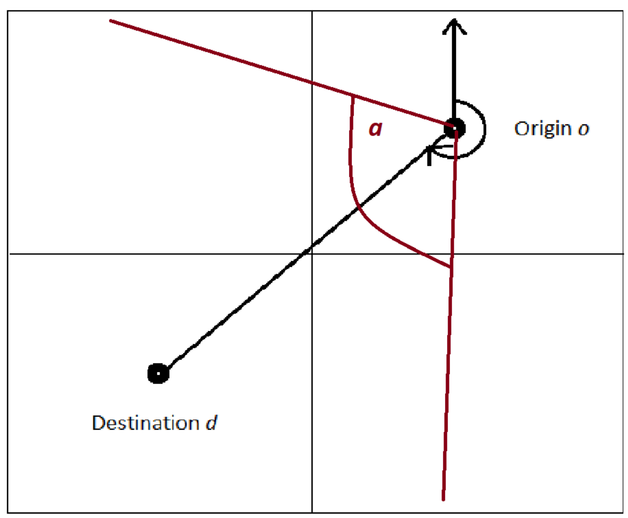


Figure 2.2: Schematic for construction of fire exposure for Econometric Transport Model ↩

Note: The pink lines on top are fractions of time during the day when the wind was blowing in that bin.

Figure 2.3: Rice area and fire counts by districts in Punjab and Haryana (2002-2011) ↩

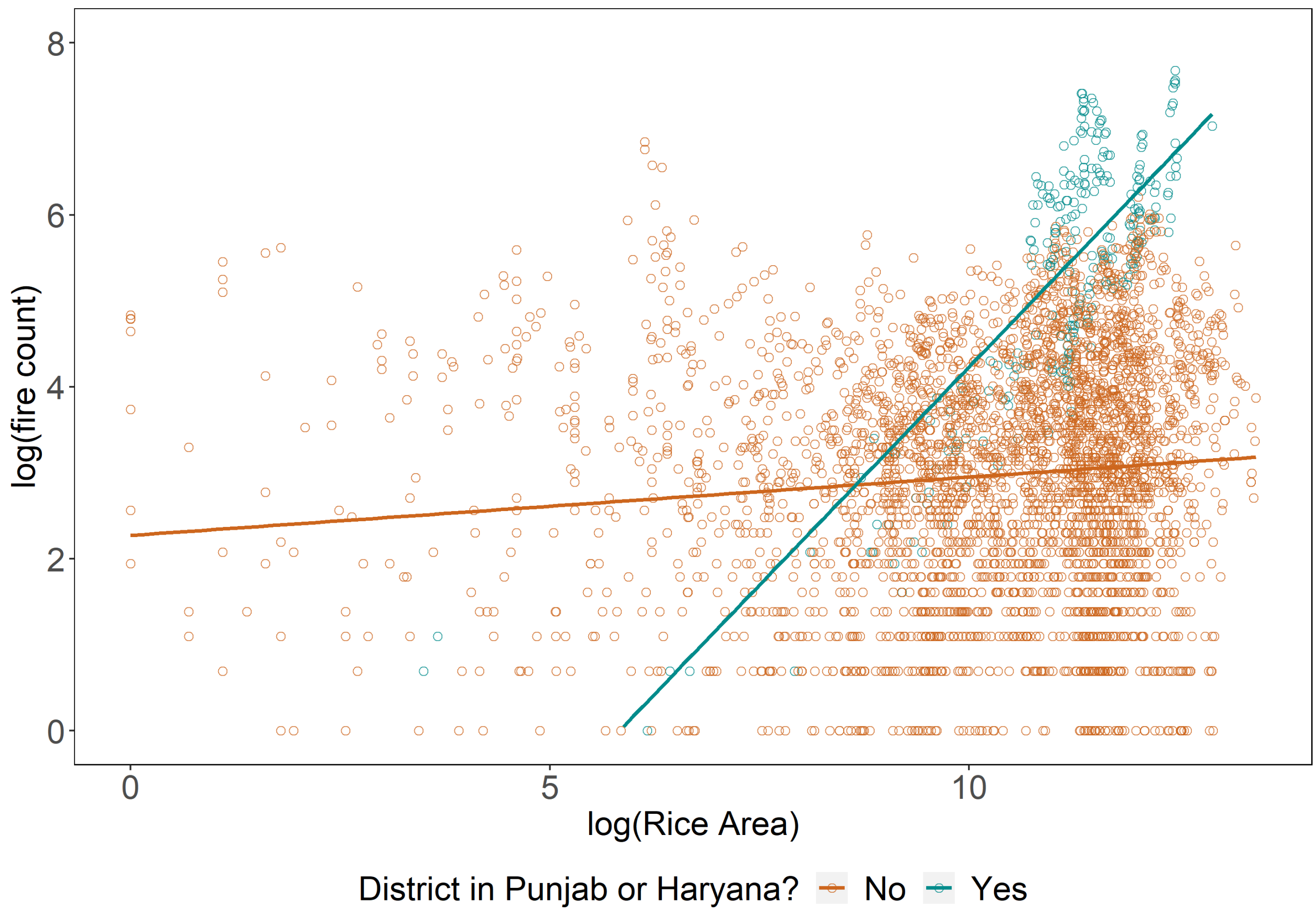
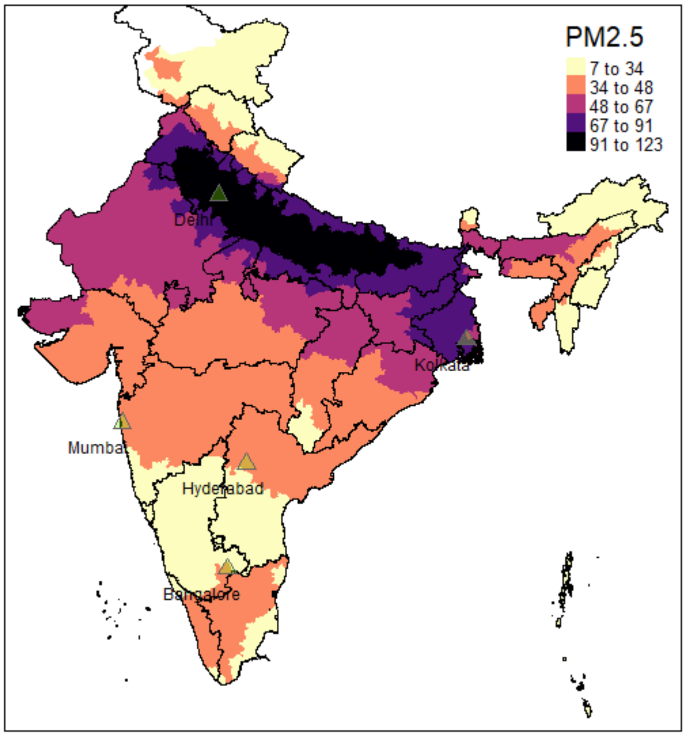


Figure 2.4: Spatial correlation between district fire exposure and PM2.5 ↩

(a) Yearly District PM2.5 Concentration (2010)



(b) Yearly District PM2.5 Exposure Ω_d (2010)

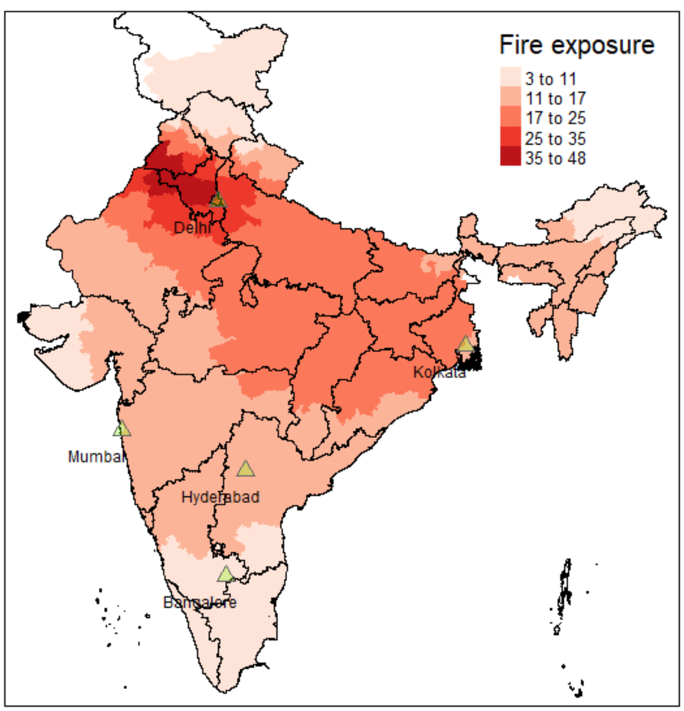
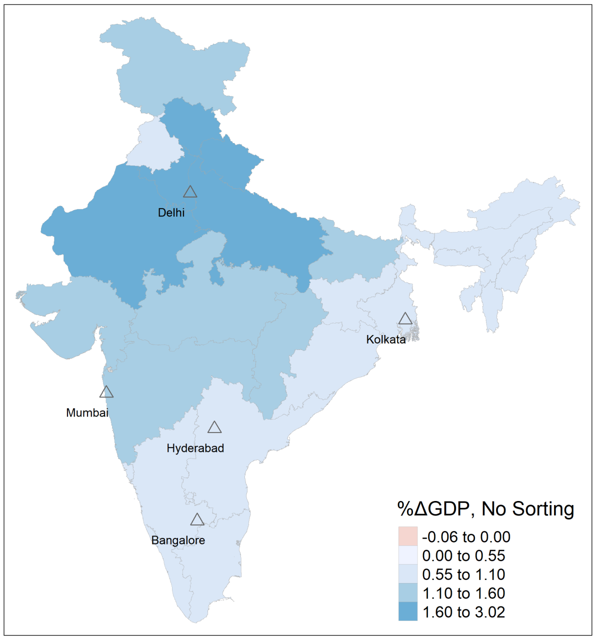
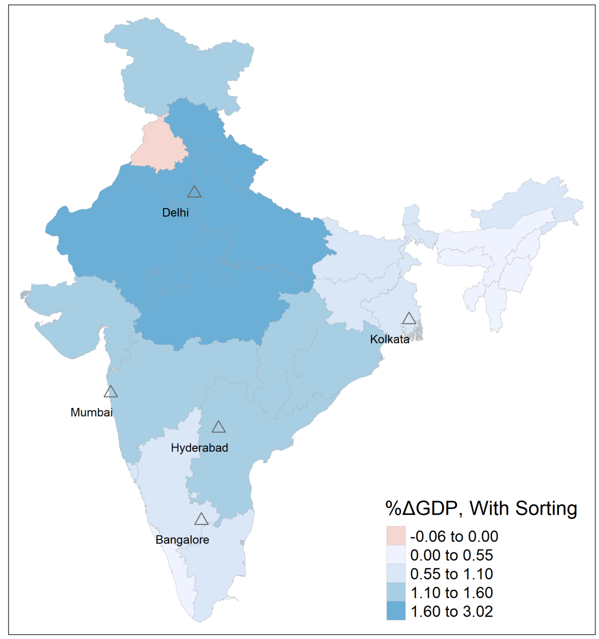


Figure 2.5: Changes in GDP and Welfare from policy to reduce fires in Punjab and Haryana ↩

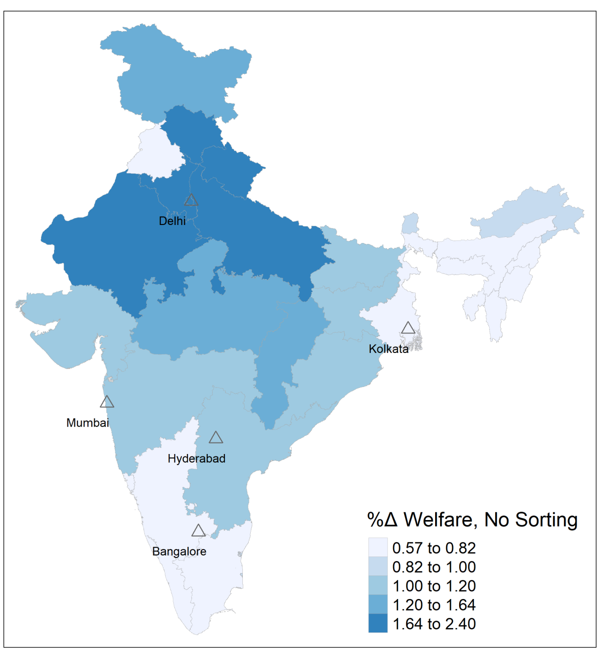
(a) Percentage Change in GDP without sorting



(b) Net % Change in GDP with sorting



(c) Percentage Change in Welfare without sorting



(d) Net % Change in Welfare with sorting

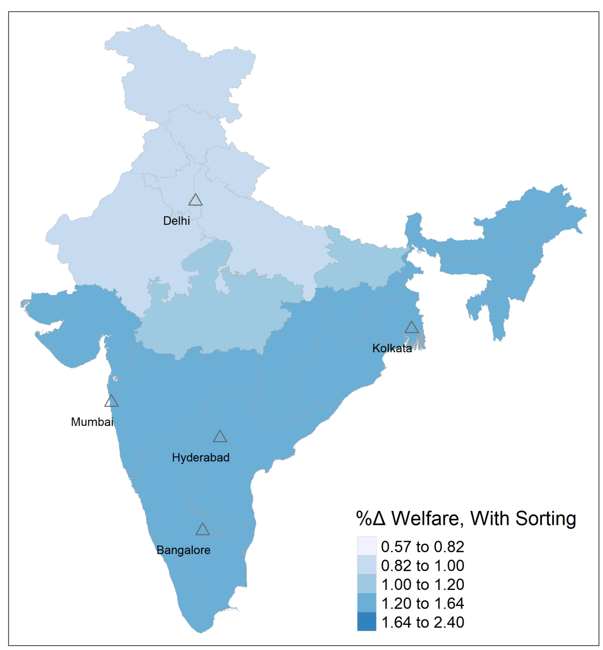
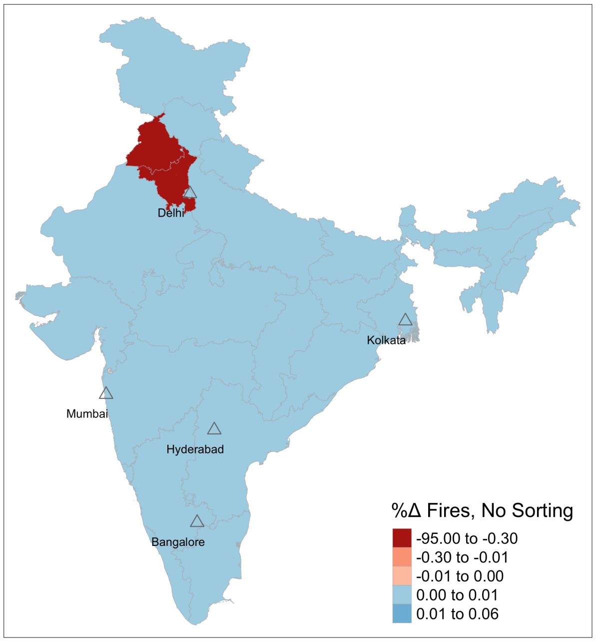
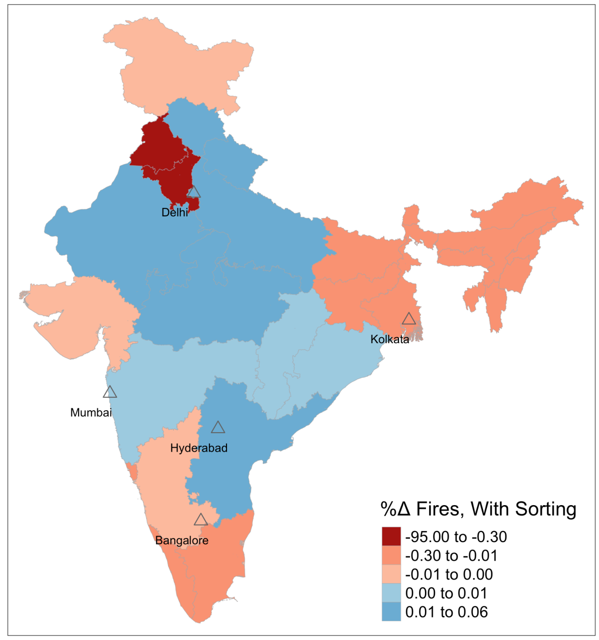


Figure 2.6: Changes in fires / fire exposure from policy to reduce fires in Punjab and Haryana ↩

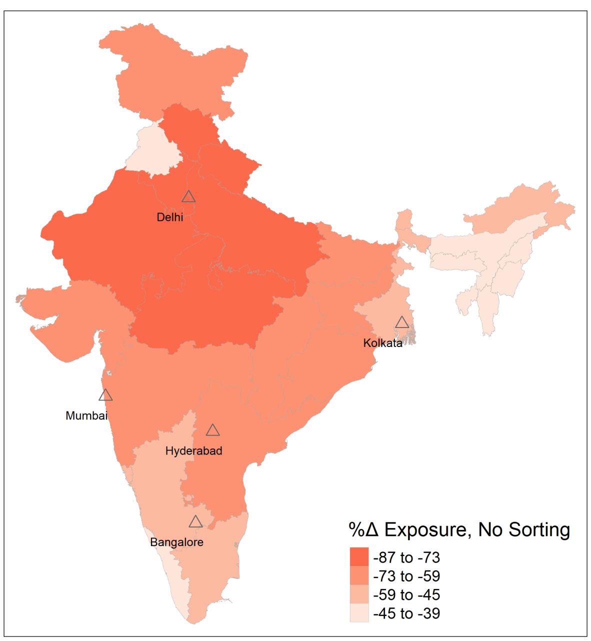
(a) % Change in Fires without sorting



(b) Net % Change in Fires with sorting



(c) % Change in Exposure without sorting



(d) Net % Change in Exposure with sorting

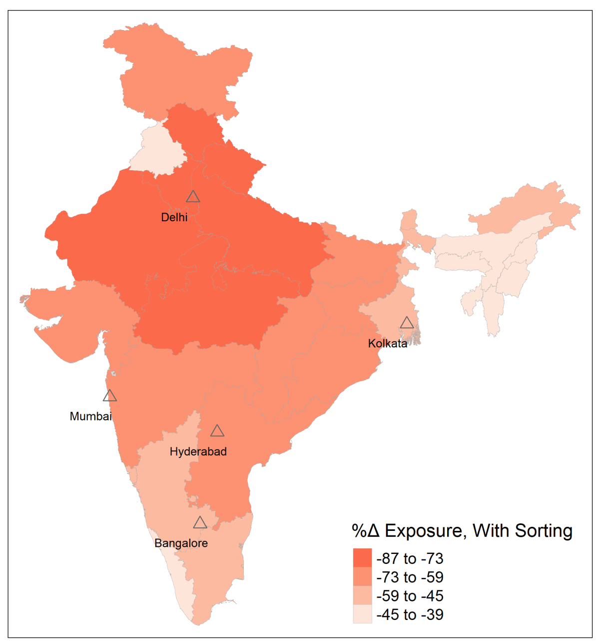
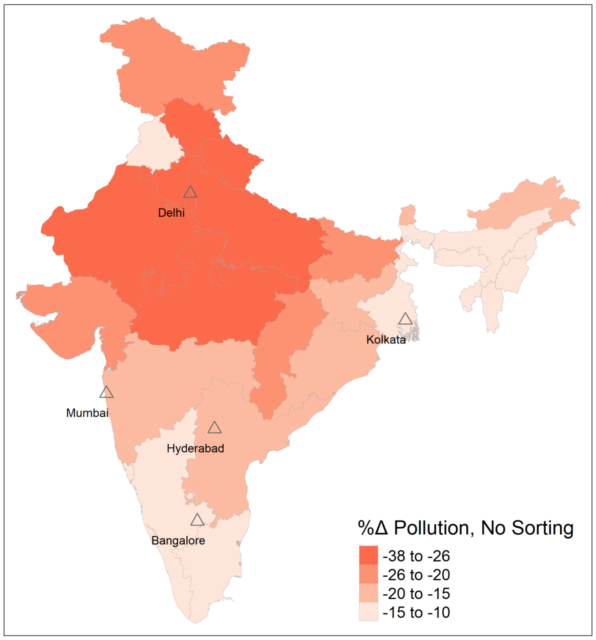
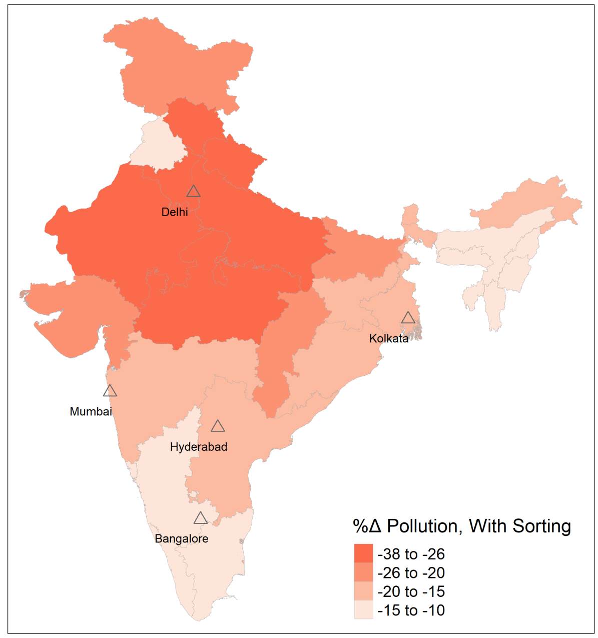


Figure 2.7: Changes in pollution and labor from policy to reduce fires in Punjab and Haryana ↩

(a) % Change in Pollution without sorting



(b) Net % Change in Pollution with sorting



(c) Net % Change in Labor with sorting

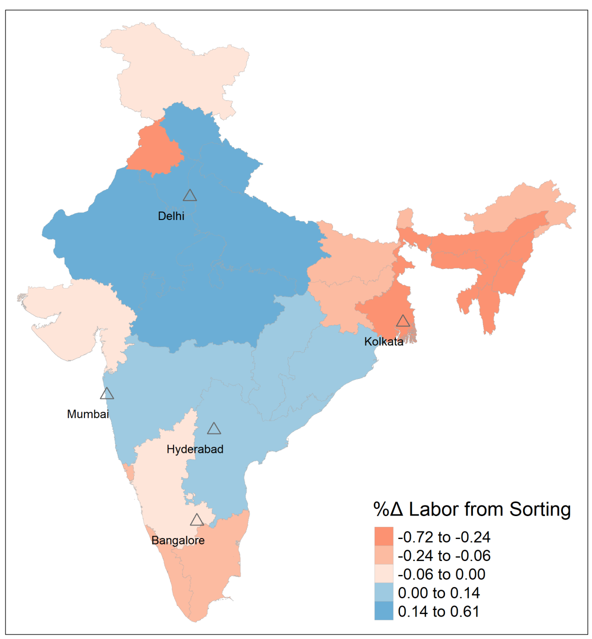


Table 2.1: Model Parameters Estimated from Data

Parameter	Description	Equation
γ	Pollution elasticity of local fire emissions	2.3
τ	Pollution elasticity of external fire exposure	2.3
δ	Emissions elasticity of labor in agriculture	2.12
ν_1	Distance elasticity of migration flows (Physical gravity)	2.14
ν_2	Language elasticity of migration flows (Cultural gravity)	2.14
λ	Pollution elasticity of migration flows	2.14
η	Income elasticity of migration flows	2.14

Notes: Estimation of parameters is done at the district level except for τ and γ which utilize pixel-level data, but close to the resolution of a district. ↩

Table 2.2: Summary Statistics

Variable	N	Mean	SD	Min	Max
Panel A: District data for estimation of fire elasticies δ, 2003-2011)					
Area under rice cultivation (ha)	5367	131595	118007	1	920015
Fire count	5367	70.65	165.02	1	2345
Panel B: Pixel data for estimation of Ag pollution elasticites γ and τ (2002-2018)					
PM2.5 (microgram/m3)	4862	50.60	23.45	5.32	137.4
Fire count	4862	127.35	403.1	0	5386
Fire Exposure (Count)	4862	16.78	8.67	1.491	69.144
Panel C: District data for estimation of migration elasticies η and λ (2010)					
Migration share	312870	0.002	0.037	0.000	0.982
Real Wage (Rs)	312870	244.445	148.149	77.028	2757
PM2.5 (microgram/m3)	312870	56.513	25.478	11.944	123
Distance (km)	312870	1032.133	572.250	0	3005
Language indicator	312870	0.735	0.441	0	1
Fire Exposure (Count)	312870	132.157	123.345	2.380	704.974
Neighbor wage (Rs)	312870	239.789	105.071	102.577	991

Notes: Summarizes the data used to estimate the main parameters of the quantitative model. Panel A presents district level data on rice area and fire counts for 495 districts that grow rice. Panel B describes pixel level fire counts, fire exposures and particulate matter data at a 1 degree resolution. Panel C describes pair-wise migration shares data across all districts from the 2011 census. The data correspond to the period of April 2010 - December 2010, with average particulate matter for calendar year 2010. ↩

Table 2.3: Effect of agricultural fires on local and external PM2.5

	Dependent variable: log(PM2.5)					
	(1)	(2)	(3)	(4)	(5)	(6)
log(Fire Count) [γ]	0.014**	0.003				
	(0.005)	(0.005)				
ihs(Fire Count) [γ]			0.015***	0.005		
			(0.005)	(0.005)		
log(1+Fire Count) [γ]					0.018***	0.006
					(0.006)	(0.006)
log(Fire Exposure) [τ]		0.226***		0.243***		0.242***
		(0.042)		(0.044)		(0.044)
Fixed Effects	District, Year	District, Year	District, Year	District, Year	District, Year	District, Year
Observations	4,530	4,530	4,862	4,862	4,862	4,862
R2	0.966	0.969	0.969	0.972	0.969	0.972
Within R2	0.011	0.116	0.012	0.128	0.014	0.128

Notes: Years 2002-2018. Estimation is done on pixels at a resolution of 1°, approximately the size of an average Indian district. Column 1 tests whether log(fire count) predicts PM2.5 by itself. Column 2 adds in log(Fire Exposure). Columns 3/4, and 5/6 do the same but with different measures of fire count to test sensitivity to dropping zeros. *p<0.1; **p<0.05; ***p<0.01. Standard errors clustered at pixel and year. ↩

Table 2.4: Estimation of fires elasticity to rice residue (δ)

	Dependent variable: log(Fire Count)
	(1)
log(Rice Area) x $\mathbb{1}(d \in PH)$	0.499***
	(0.181)
log(Rice Area) x $\mathbb{1}(d \notin PH)$	0.099**
	(0.048)
Fixed-effects	District, State X Year
Number of Districts	495
Observations	3,996
R2	0.903

Notes: Years 2003-2011. District-level regressions to estimate institutional parameter governing fire elasticity from rice residue burning. PH refers to districts in Punjab and Haryana. Estimation uses rice area as dependent variable since it is proportional to labor used in rice cultivation, and available yearly. Standard errors clustered at district and state-by-year. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 2.5: Income and Pollution elasticities of migration

	Dependent variable			
	$\log(\pi_{od})$	log(pm)	log(real wage)	$\log(\pi_{od})$
	(1)	(2)	(3)	(4)
log(PM) [$\eta\lambda$]	-0.192		-0.011	-0.824***
	(0.120)		(0.045)	(0.212)
log(Real wage) [η]	0.807***	0.002		1.51**
	(0.144)	(0.031)		(0.530)
log(Distance) [$\eta\nu_1$]	-1.78***	0.025***	0.009*	-1.79***
	(0.024)	(0.004)	(0.005)	(0.026)
$\mathbb{1}$(Different language) [$\eta\nu_2$]	-0.362***	-0.182***	0.061***	-0.566***
	(0.085)	(0.017)	(0.021)	(0.103)
log(FRP exposure)		0.709***		
		(0.036)		
Nbr distance-wtd real wage			0.240***	
			(0.059)	
Observations	312,870	312,870	312,870	312,870
R2	0.538	0.605	0.325	0.514
Estimate Type	OLS			2SLS
Birth District FE	X	X	X	X
First Stage Outcome		PM2.5	Wage	
KP First Stage Fstat		190.89	10.541	

Notes: Outcome data are pair-wise migration shares data across 615 districts from the 2011 census, corresponding to data collected during the period of April-December 2010. Other independent variables are calculated for calendar year 2010. Standard errors clustered at district of enumeration. Weather controls for district of enumeration include temperature, windspeed, cloud cover and relative humidity included. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 2.6: Estimated model quantities

Quantity	Description	Estimation
A^a	Ag productivity	Data
A^n	Non-Ag productivity	Data
$\bar{A}^a$	Fixed component of A^a	Fit eqn (2.8) to data
$\bar{A}^n$	Fixed component of A^n	Fit eqn (2.8) to data
$\bar{Z}$	Fixed pollution component	Fit eqn (4.5.1) to data
B	Fixed Amenity	FE from (2.14)
θ_d	Rice share	Data

Notes: Estimation of parameters is done at the district level to increase power unless otherwise specified. The model is estimated at the state level as sectoral GDP is not available below the state level.
↩

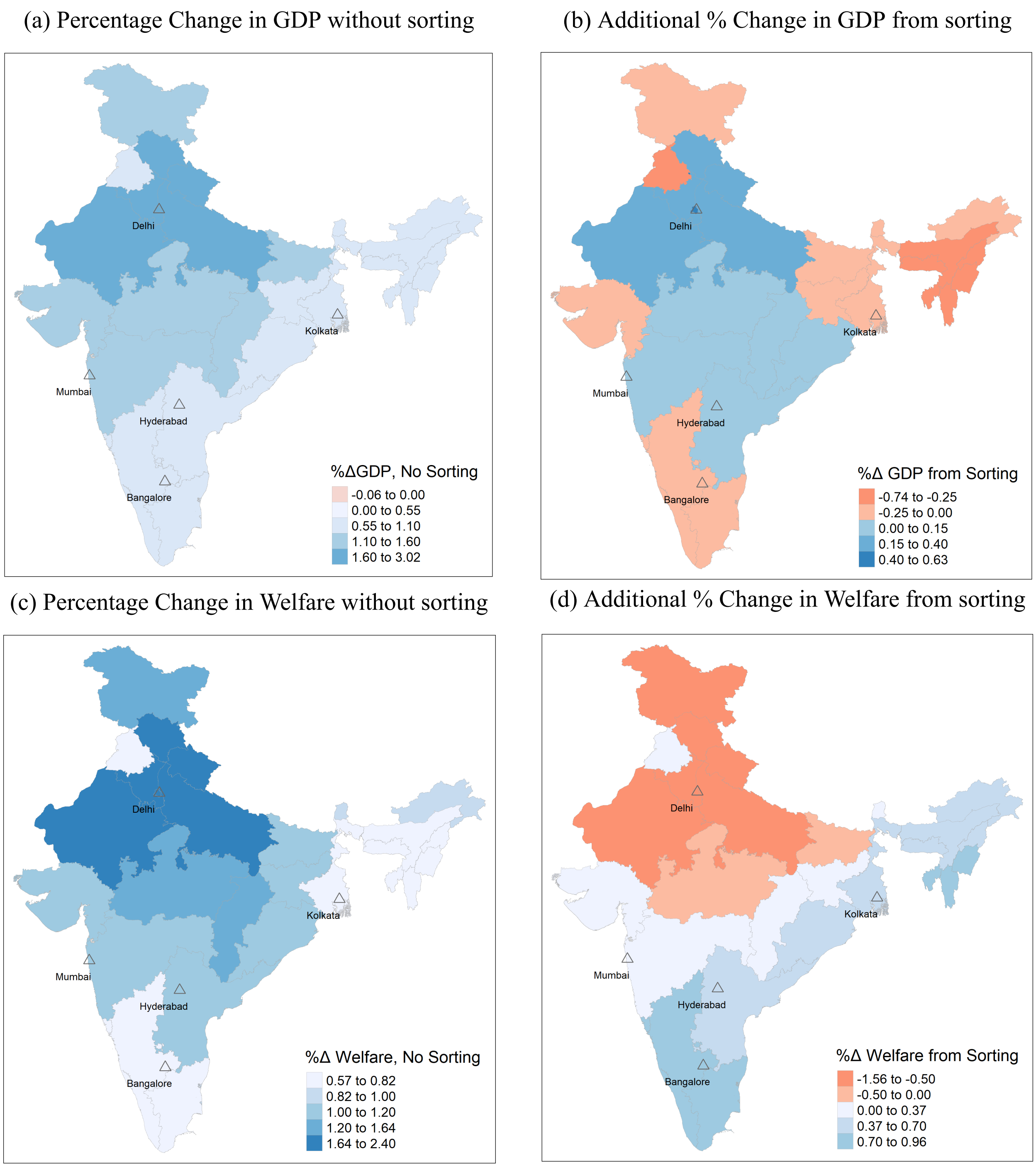
Table 2.7: Counterfactual GDP and Welfare changes

	GDP	Welfare
Net Change	1.22%	1.29%
Gain from sorting	0.01%	0.39%
Change in Gini coefficient	-0.23%	0.27%

Notes: Displays results from the policy counterfactual to reduce fire intensity in Punjab and Haryana to levels in the rest of the country. ↩

2.9 Appendix

Figure 2.A1: Additional change in GDP and Welfare from sorting with policy to reduce fires ↩



Chapter 3

The Air Pollution-linked Productivity Impacts of a Groundwater Conservation Policy

3.1 Introduction

The relationship between economic growth and the environment at various levels of development is poorly understood. The Environmental Kuznets Curve hypothesized an inverted-U shape, with economic growth worsening environmental outcomes at low levels of GDP per capita, but improving these outcomes at high levels (Grossman and Krueger 1995). Recent evidence has disputed this characterization, instead focusing on uncovering the causal factors behind various environmental outcomes at different stages of development[1] (Jayachandran 2022; Stern 2017). This paper documents that the protection of groundwater resources in India substantially increased air pollution, resulting in substantial cost to economic growth.

Aquifer depletion due to over-exploitation of groundwater is a well-documented and pressing problem in India (World Bank 2021), Aquifer depletion has docu-

[1] For a sample of urban areas across the world, Jayachandran (2022) documents that greenhouse gas emissions keep increasing with GDP per capita, lead pollution displays the EKC inverted-U pattern, air pollution (particulate matter concentrations) displays a linear and correlation while Ozone does not display any correlation at all.

mented economic costs today (Blakeslee et al. 2020) with adaptation to long-term water loss uncertain (Hagerty 2021). In order to combat particularly acute groundwater depletion, the states of Punjab and Haryana in India passed groundwater conservation laws that inadvertently increased concentrations of particulate matter less than 2.5 microns in diameter (PM2.5) locally and across inter-state boundaries.[2] India has the highest average PM2.5 concentrations in the world at 7 times the WHO standards (Greenstone 2021). The economics literature documents wide-ranging impacts of this type of air pollution, including on human health and mortality (Schlenker and Walker 2016; Deryugina et al. 2019) and labor productivity (Graff Zivin and Neidell 2012; Chang et al. 2019; Fu et al. 2021; Adhvaryu et al. 2022; Borgschulte et al. 2022) among others. This paper quantifies the short-term consequences for economic growth of increases in PM2.5 driven by the groundwater conservation laws, and summarized in the Gross Domestic Product (GDP). Thus, decisions to protect local groundwater resources in the interest of local long-term growth caused inter-state air pollution externalities with wide-ranging and immediate economic costs.

The groundwater depletion problem in Punjab and Haryana has its roots in the worldwide Green Revolution of the 1960s that allowed poor countries such as India to grow sufficient food and avoid regular famine events (Pingali 2012). The set of institutions that took root during that period in Punjab and Haryana led to the cultivation of water-intensive rice crop in these states that had not cultivated any rice before. Farmers were incentivized to pump out groundwater to irrigate paddy fields.[3] By the early 2000s, the state governments had realized the problem; but rather than incentivizing a shift away from the rice crop or instituting a marginal price for groundwater[4], these states decided to force farmers to push back the dates

[2] Other pollutants such as Ozone and PM10 may also be correlated with increases in PM2.5.

[3] Rice is synonymous with paddy, with interchangeable references to paddy or rice fields.

[4] The marginal price of groundwater extraction continues to be close to zero, since electricity for pumps is subsidized with flat tariffs rather than marginal pricing, and any outstanding farm electricity bills are rarely paid back to state distribution companies.

of rice transplantation[5] from mid-May to mid-June, hoping that the arrival of the monsoon by late June would reduce dependence on groundwater to grow rice.

Even before these groundwater conservation laws were passed in 2009, farmers in Punjab and Haryana would set fire to their fields after the rice harvest in October in order to prepare the same fields for planting the staple wheat crop. These fires clear out leftover residue after rice harvest that come in the way of planting wheat seeds, and have become popular as the cheapest method to get rid of this residue. Although agricultural fires are nominally illegal, enforcement is rare, with an average expected fine in Haryana of 0.75 USD while the average marginal cost to clear the field without fires is 50 USD (Behrer 2019; Lohan et al. 2018). By forcing a shift in the transplantation dates to early June in order to conserve groundwater, these states also shifted rice cultivation dates from October into November. Since any delays in planting of wheat crops can reduce yields (McDonald et al. 2022), farmers had further incentives to utilize fires to get rid of the rice residue after a later harvest due to the laws. These factors had the unintended consequence of shifting the peak of agricultural fires from October into November, when the onset of winter brings lower wind speeds and temperatures that slow the dispersion of particulate matter (Vallero 2014).

I utilize a two-way fixed effects design with information on the timing of the groundwater laws to document that the laws increased November fire count by 54% and the measure of biomass burnt by 72% in districts of Punjab and Haryana. Small anticipation effects imply that these may be slight underestimates. At the same time, October fire count and measure of biomass burnt decreased by 58% and 57% respectively. Since winter fire activity is predominantly concentrated in these two months, these results document the shift in monthly fire patterns toward early winter. Next, I develop a novel method to capture the effect of the increase in November fire activity on annual PM2.5 levels. November fire exposure strongly

[5] This process of moving seedlings grown in nurseries into fields reduces weed removal and produces higher yields. More details at http://www.knowledgebank.irri.org/training/fact-sheets/crop-establishment/manual-transplanting

affects annual PM2.5 levels across India, with changes in fire exposure explaining 4.2% of annual deviations in within-district PM2.5, compared to 16% explained by local weather.

Next, I analyze the effect of PM2.5 levels on district GDP using newly available panel data for 530 Indian districts between 2007 and 2013. In order to account for the non-stationary nature of the GDP data[6], I utilize district-specific time trends as well as a first differences approach. The latter performs better with strongly non-stationary data series and is commonly used in macroeconomic analysis of GDP data (Wooldridge 2010). Identification of the causal effect of PM2.5 on district GDP relies on yearly deviations in PM2.5 being plausibly exogenous, controlling for year and district fixed effects as well as time trends. However, causality may yet run from GDP to PM2.5, with larger yearly deviations in pollution systematically being a result of higher economic activity in that district in the given year. To tackle this endogeneity concern, I instrument for PM2.5 using the novel variable linking all upwind fires to local PM2.5 concentrations. With this instrument, I show that a 1% increase in PM2.5 levels reduces GDP by 0.18% in Indian districts.

With these causal estimates in hand, I calculate the effect of the increase in November fires due to the groundwater laws on net GDP in India. Districts that are closer to and downwind of districts in Punjab and Haryana see a larger increase in particulate matter concentrations, and therefore reductions in GDP. I calculate an estimate of yearly net GDP losses using the three estimated elasticities: the increase in November fire strength due to the laws, the increase in downwind PM2.5 due to higher November fire exposure, and the reduction in GDP from an increase in PM2.5 levels. This leads to an estimated yearly loss of 0.125% of national GDP due to the groundwater laws. For comparison, the share of yearly expenditure as a percentage of GDP on the National Rural Employment Guarantee Scheme (NREGS), a flagship welfare program, is about 0.25%. The loss figure of 0.125% is also an underestimate of the overall economic costs associated with the increased pllution

[6] Average yearly district GDP growth rate in this period was 7%.

due to these laws, since it does not account for increased infant and old-age mortality, as well unaccounted expenditures on health, lost schooling years etc. that are not monetized into GDP.

This paper contributes to the literature on how institutions affect environmental outcomes in developing countries. More stringent regulation (Burgess et al. 2019), use of technology (Assunção et al. 2022) and higher resource allocation to monitoring and enforcement (Duflo et al. 2018) can improve environmental outcomes in developing countries. Weak state capacity impedes the implementation of regulations on the books that prohibit the use of fires in agriculture in Punjab and Haryana, leading to large economic costs.

Another explanation may be that while the groundwater externality is localized to the two states, the air pollution externality is an inter-state phenomenon. Lipscomb and Mobarak (2017) show that decentralization of regulatory authority in Brazilian municipalities leads to larger water pollution externalities across border. Kahn et al. (2015) document that providing promotion incentives to reduce some water pollutants to local officials reduces their externality on downstream neighbors. While the states of Punjab and Haryana are able to successfully implement one set of laws intended to conserve local groundwater, they are unable (or unwilling) to implement regulations on fires, which cause downwind externalities on top of local ones. This result suggests that designing regulatory institutions for environmental protection at the appropriate level is important in determining the outcomes of regulation.

Finally, this paper also relates to the literature on second-best institutions in developing countries. Rodrik (2008) argue that institutional design in developing countries with multiple distortions should not insist on the first-best, since the desired outcome can be achieved at lower cost through second-best practices. The textbook, first-best solution to the groundwater externality in Punjab and Haryana is marginal pricing of groundwater. But this is very unlikely to occur given the political power of farmer lobbies in these two states. Unfortunately, the second-best solution to utilize the monsoon rains for rice cultivation backfired by exacerbating the air pollution externality.

The rest of the paper is structured as follows. Section 2 describes the data while section 3 presents the context and motivates the QSE model by estimating the extent of pollution externalities from agricultural fires. Section 4 presents the model of general equilibrium while section 5 describes the estimation of the parameters governing equilibrium. Section 6 conducts counterfactuals and section 7 concludes.

3.2 Background

India is the largest user of groundwater in the world; but with almost 20% of the world's population, it only has about 4% of the world's freshwater resources (World Bank 2021). The resulting overuse of groundwater to meet population needs has caused rapid aquifer depletion and led to an urgent environmental crisis, particularly affecting the alluvial plains of North-Western India. This section first discusses the factors behind this depletion in the North-Western states of Punjab and Haryana, leading to the passage of a set of groundwater conservation laws in 2009. I then discuss how these laws unintentionally pushed agricultural fires into early winter, when their impact on air pollution is exacerbated due to meteorological conditions.

3.2.1 Groundwater conservation laws in Punjab and Haryana

Until the advent of the so-called Green Revolution of the 1960s that raised agricultural productivity dramatically across India and much of the poor world (Pingali 2012), North-Western India was a primarily wheat-growing region with little rice consumption or production locally. One of the institutional innovations of the Green Revolution in these states was the provision of large subsidies for tubewells and borewells. Individual farmers could now access shallow groundwater to irrigate fields even if they did not have access to the large, pre-existing canals systems built by the colonial British empire. Over time, modern pumps running on electricity were combined with practically zero tariffs to farmers so that they could irrigate their fields at minimum cost.

This newfound access to groundwater allowed farmers to diversify their crop portfolio by allowing the cultivation of the highly water-intensive rice crop during the "Kharif" or monsoon season (June-October). The wheat crop is cultivated during the "Rabi" or winter season, when the lower temperatures and plenty of sunshine provide perfect weather conditions for growth (Kataki et al. 2001); planting happens in early winter and harvest in early spring.

The state of Punjab contributed less than 1% of India's rice in 1961; by the late 1990s this figure was up to 10%; absolute rice output across India rose from 11 million tonnes to 75 million tonnes in this period, underlining the massive increase in rice cultivation in Punjab (Subramanian 2017). Similar trends in rice cultivation were seen in Haryana. This fundamental change in the cropping patterns of the region exacerbated the depletion of groundwater resources, since the paddy fields were flooded primarily using groundwater, pumped out before the annual monsoon reached Punjab and Haryana. Taken together, the unregulated exploitation of groundwater had led to an acute water crisis by the early 2000s, although concerns about excessive extraction almost 1.5 times the natural recharge rate had been expressed by agricultural scientists and government committees going back to the 1980s (Singh 2009).

Despite the alarm expressed by various stakeholders, the state governments largely ignored the problem until the early 2000s. When asked about these concerns, the then-Chief minister of Punjab, Prakash Singh Badal, is quoted in the media as saying, *"The problem is not as acute as is being projected. It is a theoretical evaluation and there is no truth in it"* (Down To Earth 1999). The political economy of both states, but particularly of Punjab, centers around medium and large sized farmers who receive a range of state subsidies that incentivizes rice and wheat cultivation. Apart from the Green Revolution era technological subsidies for higher-yielding seeds, fertilizers and pesticides, tubewells and electric pumps, provision of cheap electricity is also important in explaining groundwater levels (Ryan and Sudarshan 2020). Often these dues are not paid to state distribution companies at all, resulting in lack of investment in the power grid (World Bank 2018). But, most importantly, the procurement of wheat and rice crop by the state governments of Punjab and Haryana at so-called Minimum Support Prices that distort market signals (Parikh et al. 2003) precludes farmers from switching to other crops with higher price and yield risks.

The practice of transplantation of rice before mid-June was thought to be particularly cause too much reliance on groundwater (Singh 2009). In response, sections of the state bureaucracy had made efforts starting in the early 2000s to shift the transplantation of rice closer to the monsoon, since this was thought to ease the strain on groundwater use. The two governments took executive action through ordinances in 2008[7] to extend the practice of delaying rice transplantation statewide. Given the generally favorable response to this ordinance, the legislatures of Punjab and Haryana separately ratified the Preservation of Subsoil Water Acts of 2009 ("laws" from now on) in an effort to conserve groundwater.

These laws prohibited early transplanting of rice before the monsoon in an attempt to reduce groundwater usage for irrigation. Much of the rice transplantation would occur in the peak of summer during May when evapotranspiration (water loss from plants as well as soils and water bodies) is high. These laws specified June 10 as the earliest transplantation date, and it was shifted further to June 20 later[8]. When planting rice in May, farmers were solely dependent on groundwater reserves for rice growth; moving transplantation to June allowed rice growth to depend more

[7] These do not have the same power in Indian law as a statute and cannot be renewed beyond a few months.

[8] The Indian Met Department (IMD) sets out July 1 as the expected date of onset of the monsoon in North-Western India. Details here

on monsoon rainfall. This was expected to lead to a lower rate of groundwater extraction.[9]

3.2.2 Increase in agricultural fires due to shifting of rice crop harvest

The primary use of fires in Indian agriculture today is to clear the field of leftover residue from harvesting a crop, before sowing and planting the next crop (Shyamsundar et al. 2019); this differs from slash-and-burn agriculture that is practiced in parts of Africa and Indonesia (Andini et al. 2018). Figure 3.1 shows that agricultural fires are concentrated in the states of Punjab and Haryana, which are also characterized by a Rice-Wheat crop system.

In the Rice-Wheat system of Punjab and Haryana, fires are used to clear rice residue before the planting of the wheat crop on the same land, since rice residue comes in the way of planting wheat. This practice dates back at least to the 1990s. The earliest observations of fires from the NASA FIRMS database (described in the next section) starting in 2002 clearly demonstrate that North-western India already had a disproportionate share of fires in Indian agriculture.

The delay in rice transplantation due to the laws also pushed back harvest dates. The resulting delay in rice harvest from mid October to late October and November meant that farmers had fewer days between rice harvest and wheat plantation. Any delays in wheat plantation beyond the first two weeks of winter reduces yields substantially (McDonald et al. 2022). Therefore, the law had the unintended consequence of increasing the intensity of agricultural fires in November, when slower winds and lower temperatures tend to worsen downwind air quality.

[9] Groundwater recharge is typically a slow-moving process that takes place over a longer period than the period of study here. I plan to conduct an assessment of the change in groundwater levels to the present day due the policy in the future

3.3 Data and Measurement

3.3.1 Agricultural fires

The burning of residue from crop harvest is referred to as agricultural fires. To analyse the impact of the groundwater conservation laws on the monthly pattern of fires, the ideal data would include precise location of each individual fire set for the purpose of burning crop residue. But there are no representative ground-level observations of this phenomenon. To overcome this challenge, I utilize the Fire Information for Resource Management System (FIRMS) product from the National Aeronautics and Space Administration (NASA) agency of the United States that is widely used to identify terrestrial fires. This product provides information on daily fires detected at latitude/longitude level across the world and has been recently used to analyze agricultural fires in the economics literature (Behrer 2019).

FIRMS provides a few related products: Near-Real Time (NRT) fires using the MODIS instrument aboard Terra and Aqua satellites, standard product from the same instrument but with a 2-3 month lag and another NRT product using the VIIRS instrument from the Suomi-NPP and NOAA-20 satellites. The main difference between the first two and the third is the resolution of the data. MODIS products are at 1 km resolution and are available from 2000 (more reliable from 2002 when Aqua satellite was launched) whereas VIIRS products are at 375 m but only available from 2012. The primary analysis utilizes the MODIS standard product which differs from the NRT data in that corrections are made to the imprecise location of the Aqua satellite in the NRT data. Imagery data from Aqua and Terra satellites is available at least four times daily for each pixel on Earth and is processed using a NASA algorithm isolate a ground-level fire signal from other signals such as solar flares.

I combine this data with information on land use from the European Space Agency

Climate Change Initiative's land cover map (version 2.07).[10] This allows the subset of fires that is found on agricultural land to be separated from natural forest fires since this paper is interested in agricultural fires. I aggregate and resample the land cover data which is at a resolution of 300 m to the fire data grid (at 1 km resolution), with an indicator for agricultural land use as the main output from this process. All fires are then masked based on this indicator variable to find the subset of agricultural fires.

3.3.2 Air quality

An important consideration for air quality data is complete geographical coverage. Whereas ground-level monitoring station coverage in India is extremely sparse (Greenstone and Hanna 2014), satellite imagery-based products provide complete coverage. Secondly, ground-level observations may be susceptible to manipulation (Greenstone et al. 2022; Ghanem and Zhang 2014). Therefore, the main source of data on air quality is Hammer et al. (2020), a gridded reanalysis product of global surface PM2.5 concentrations at a resolution of 0.01° that should be much less susceptible to such manipulation. This product combines satellite imagery data on Aerosol Optical Depth with state-of-the-art chemical transport models, and calibrates the output to global ground-based observations. It is easy to aggregate the gridded product to the necessary resolution for analysis at pixel, city or district level. Forthcoming sections will detail the aggregation procedure for each analysis.

3.3.3 GDP data

To estimate the impacts of PM2.5 on GDP, we would like data at the most granular sub-national level possible. While satellite-based data on PM2.5 levels are available at a 1x1 km grid, output data are rarely available at such sub-national scales. Fortunately, GDP measures at the district level in India between 2007-

[10] Data is available at https://cds.climate.copernicus.eu/cdsapp/#!/dataset/satellite-land-cover

2013 have recently been compiled by the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in their District Level Database (DLD).[11] I clean and combine these data with other district-level data using district identifiers from ICRISAT and the Census of India, 2011.

3.3.4 Meteorological data

Hourly wind data are used to construct exposure to agricultural fires for every origin-destination pixel pair. Details of the methodology follow in the section 3 below. The source of these wind data is the European Center for Medium Range Weather Forecasting (ECMWF) ERA5 family of global gridded reanalysis datasets.[12]. Reanalysis data combine ground-level observations and satellite data with Chemical Transport Models that represent physical and chemical processes in the atmosphere to produce reliable and complete coverage for the world. Since ground-level observations are particularly sparse in developing countries these reanalysis data are widely used in the literature on climate and air pollution in Economics (Auffhammer et al. 2013) Hourly wind speed and direction data are taken from the ERA5-Land hourly dataset which is available at a resolution of 0.1°. These are combined with daily agricultural fires at the pixel level to construct the fire exposure variable. Apart from being used to quantify the contribution of distant residue burning on local air pollution, fire exposure also is an instrument for pollution at the city and district level in estimation of certain elasticities. Finally, I also construct temporal averages for weather variables including rainfall, temperature and relative humidity from this dataset to be used as controls in the regression analysis.

[11] http://data.icrisat.org/dld/src/crops.html
[12] Data available at https://cds.climate.copernicus.eu

3.3.5 Constructing instrument for air quality using agricultural fires

Since the laws push fires in Punjab and Haryana into November when there used to be very few fires earlier, I only consider the month of November in constructing this instrument. I capture the contribution of daily agricultural fires F_{ot} from 1x1 degree origin pixel o on air quality in destination pixel d, at a distance of $dist_{od}$[13], in the month of November, as follows

$$\omega_{od} = \left(\sum_{t \in Nov} \frac{windfrac_{odt} * F_{ot}}{dist_{od}} \right) \tag{3.1}$$

$windfrac_{odt}$ is the daily average fraction of time that the wind at o blows towards d on day t. In order to calculate $windfrac_{odt}$, I start by assigning each hourly wind observation in o on day t into one of 36 bins of 10 degree span each, based on the wind direction that hour (true north is 0 degree as in the figure). I then construct the wind speed-weighted fraction of time the wind was blowing in each of these 36 bins by aggregating hourly observations for day t. $windfrac_{odt}$ is then calculated by summing up wind fraction for the bins which are in the direction of d from o as shown in figure 3.3.

I construct this instrument for various distances between origin and destination districts.[14] These are increased sequentially so that the distance that maximizes power in predicting PM2.5 can be selected (Details in the next section of this estimation). Yearly variation in the instrument is driven by two factors: (i) changes in the temporal distribution of fires at origin and (ii) changes in the daily wind patterns at origin during November.

Table 3.1 summarizes the main variables used in the analysis.

[13] The distance elasticity is assumed to be -1 but will be estimated using NLS

[14] Distances are calculated using district centroids.

3.4 Research Design

3.4.1 Effect of policy on local fires

I utilize a difference-in-differences research design with fixed effects to test how the groundwater conservation laws shifted the monthly pattern of fires. The outcome variables in each district-month-year period from 2002 to 2020 are the count of fires and the total strength of these fires as measured by the fire radiative power. These are aggregated to the district-level to reflect the administrative unit at which state policy is implemented in India. I estimate a Poisson fixed effects model to recover the coefficient of interest, assuming the standard exponential link function (Behrer 2019; Ranson 2014) for the count or measure of biomass burnt F_{dmy} in district *d*, month *m* and year *y*. The conditional expectation function given regressors $\mathbf{X}_{dmy}$ is as follows

$$\mathbf{E}[F_{dmy}|\mathbf{X}_{dmy}] = exp(\sum_{m\in[1,12]} \delta_m \, D_{dmy} + \alpha_d + \tau_{sm} * Y_y) \tag{3.2}$$

where the RHS inside the exponential function contains $\mathbf{X}_{dmy}$. Since the laws came into force at the state-level in 2009, the treatment indicator D_{dmy} turns on for district-months in Punjab and Haryana in and after 2009. District fixed effects D_d control for unobserved determinants of fires that do not change over time. Comparison of fires within state-by-month cells (τ_{sy}) flexibly controls for other within-state determinants of fire seasonality such as different crop calendars, crop mixes etc. that do not change over time. Year fixed effect Y_y controls for any common trends across the country (such as the country-wide increase in fires driven by the Mahatma Gandhi National Employment Guarantee Scheme or NREGS documented by Behrer (2019)).

The count nature of the data and the nontrivial presence of zeros in the count data motivate the use of a Poisson model. A log transform of F_{dmy} would create bias in a linear model estimation whereas an inverse hyperbolic sine transform makes the interpretation of the elasticity slightly more complicated (Bellemare and Wichman 2020). Further, the Poisson FE model only requires that the conditional expectation function be specified correctly for consistent estimation of the parameters (Wooldridge 2010). It produces unbiased estimates of the coefficients even if the fire data do not match the Poisson distributional assumptions (Wooldridge 1999a, 1999b; Lin and Wooldridge 2019). This is not true of other models that are used to handle count data such as negative binomial (Blackburn 2015). I estimate this model using quasi-maximum likelihood method through the *fixest* package in R (Berge et al. 2022).[15]

Taking log of (3.2) yields the following

$$log(\mathbf{E}[F_{dmy}|\mathbf{X}_d my]) = \sum_{m \in [1,12]} \delta_m \, D_{dmy} + \alpha_d + \tau_{my} \tag{3.3}$$

Therefore the coefficients of interest δ_m give the monthly elasticity of fire count to the policy. As with any difference-in-differences design, the main identifying assumption for the δ_ms is that trends in monthly fires would be similar between treatment and control districts in the absence of the policy change. I discuss this assumption in more details in the results section. Standard errors are clustered two ways at the district and state-by-year level to account for the district-level autocorrelation as well as implementation at the state-by-year level.

3.4.2 Effect of PM2.5 on GDP

This section describes the estimation of the causal impact of higher PM2.5 levels on district GDP in India. I build up to an instrumental variables strategy for PM2.5 that allows the quantification of impact of the groundwater laws on downwind

[15] The Poisson model can be used with non-integer data such as the measure of biomass burnt as well, and the strengths of the Poisson over other model when the data have nontrivial presence of zeros also holds (Wooldridge 2010)

GDP. Before describing this IV strategy, equation (3.4) presents an OLS regression model of the effect of PM2.5 on district GDP.

$$log(GDP_{dy}) = \beta\, log(PM_{dy}) + \gamma\, Weather_{dy} + g_d * t + \alpha_d + Y_y + \epsilon_{dy} \quad (3.4)$$

$$\begin{aligned} Weather_{dy} = \{&Temp_{dy},\ Temp\_squared_{dy},\ Rain_{dy},\ Rain\_squared_{dy},\\ &Relative\_Humidity_{dy},\ Surface\_Pressure_{dy},\ Windspeed_{dy}\}\end{aligned}$$

The quantity of interest β is the percentage reduction in GDP for a 1% increase in PM2.5 levels. This model contains district and year fixed effects D_d and Y_y respectively, which control for fixed factors that raise productivity or increase pollution as well as account for any common macroeconomic shocks. Weather variables such as temperature, rainfall, humidity and wind speed are known to affect PM2.5 (Dechezleprêtre et al. 2019; Bondy et al. 2020). Therefore, I control for yearly average weather that could determine the level of pollution from given emissions.

The main residual concern with identification of β in this model is that deviations of GDP and PM can be jointly determined. Higher economic activity in a given year can itself cause an increase in PM that year by increasing emissions. At the same time, higher deviation in PM can stunt GDP growth that year through channels such as increased worker morbidity and lower labor productivity

I adopt three approaches to tackle this endogeneity issue. Figure 3.4 shows that GDP exhibits strong growth in this period and therefore is not stationary; between 2007-2013, average Indian GDP growth rate was 7%. First, I fit a district-specific linear time trend $g_d * t$ in GDP. The time trend will capture district-specific factors that cause constant GDP growth, leaving only deviations from the trend line in the outcome. This approach can also help reduce omitted variables bias (OVB) that jointly determines both GDP and PM2.5 (eg. demand shocks that affect certain districts). Such OVB can cause the causal chain to run from GDP to PM2.5, leading to reverse causality that biases the estimate upwards, since an increase in economic activity increases PM2.5 levels.

Secondly, I also conduct analysis using first differences (FD) that is the preferred over fixed effects to deal with non-stationary, autocorrelated data series in both outcome and explanatory variables. Further, an FD specification that also includes a fixed effect allows for a district-specific linear growth rate g_d in the outcome. The FD approach is commonly used in the macroeconomic literature to deal with serial correlation in aggregated GDP data. Equation (3.5) specifies the regression framework for the FD model.

$$\Delta log(GDP_{dy}) = \beta\,\Delta log(PM_{dy}) + \gamma\,\Delta Weather_{dy} + g_d + \Delta Y_y + \Delta\epsilon_{dy} \quad (3.5)$$

This specification examines how the growth rate of PM2.5 affects the growth rate of GDP, controlling for year-on-year changes in weather and common macroeconomic conditions. The district fixed effect g_d captures the constant growth rate of GDP in these districts. But even with the FD design, there may still be some OVB in the leftover variation, leading to reverse causality that biases the results upwards.

The third approach utilizes the fact that the stock of pollution in a district is partially due to sources outside the district, notably agricultural fires in this instance. Using the instrument for PM2.5 defined in the previous section with both the panel and FD specifications allows us to tackle the reverse causality challenge. Fires in the winter are much worse for downwind PM due to meteorological conditions that favor longer suspension and entrapment of particulate matter in the lower atmosphere of downwind districts. I also construct the same instrument with other months separately and together for the whole year, and test the hypothesis that fires in the winter are worse for PM2.5 in fire-exposed downwind districts. Before discussing the results on the effect of PM2.5 on district GDP in section 3.5.3, I discuss the results of this first stage in section 3.5.2, based on specifications in equations 3.4 and 3.5.

Inference would ideally be conducted using Conley spatial standard errors with arbitrary autocorrelation at district level, given the spatial autocorrelation in PM2.5 and GDP levels. However, I am unable to implement these standard errors for a panel data model with instrumental variables.[16] Instead I cluster standard errors at the region-year level, where regions are groups of contiguous districts that share similarities in economic and geographic fundamentals such as level of development, soil types, weather etc.[17] In this case, the dependence of PM2.5 is fundamentally spatial, and not administrative. In order to test whether clustering at the level of the region is appropriate, I plan to conduct inference using a wild cluster bootstrap later. I also plan to write code to calculate Conley standard errors in a panel IV setting.

3.5 Results

3.5.1 Effect of groundwater laws on monthly fire patterns

To begin the results section, I refer to figure 3.2 that shows some growth in the fire count and fire strength for November occurring just before the laws were passed in 2009, with a stronger trend upwards after the passage of the law, before stabilizing by 2015 or so. This suggests some anticipation effects in November, since the count of fires is trending up 2-3 years before the policy came into effect. These anticipation effects can be attributed to the informal implementation of the policy before 2009 that is discussed in section 3.2.1. This may have driven the shifts in fire patterns by slightly delaying the cultivation dates before 2009, with formal

[16] The R package **lfe** provides the command **felm** that implements Conley SEs with panel data; however it does not produce these SEs with panel IV estimation.

[17] There are 530 districts and 96 regions in the sample, so that there are 5.5 districts on average in each region. Each district had an area of approximately 100 sq km, on average.

implementation inducing a larger shift. The lack of pre-trends on October fires combined with a downward shift after 2009 supports this view. Therefore, the effect for November my be an underestimate.[18]

Table 3.2 presents estimates of the causal effect of these laws on monthly fires, based on 3.2. Columns 1 and 3 provide the mean number of fires and measure of biomass burnt in Punjab and Haryana, before the passage of these laws. Those columns show peaks of fires during the months of April, October and November. Fires in the latter two months are used to clear the monsoon season rice residue, as described earlier. Fires in April are used to clear the wheat crop residue after the harvest is done. The time pressure of needing to be rid of the rice residue before wheat plantation that leads to the fires after monsoon rice harvest does not arise after the wheat harvest. Yet we see substantial fire activity in April. This wheat residue burning practice may have come about due to habit formation from setting fires to the rice residue. However, it is less troublesome for downwind pollution than are fires during early winter, since meteorological conditions in April do not favor suspension of particulate matter over the plains of North India.

Turning to the results in columns 1 and 3 of table 3.2 now, the main result is that the laws increase the log of expected fire count in November by 0.43, and log of expected fire strength by 0.54. Estimates for October are negative, providing evidence that the laws probably succeeded in pushing rice cultivation by a few weeks to a month, and therefore peak fire activity into November. Estimates for the other months except June and July are negative. For the months from December to May, this would suggest a domino effect of the later rice cultivation on other crop burning, since the entire crop calendar gets pushed back. The spring season fire peak (from the wheat harvest) that used to happen in April and May seems to shift slightly toward June and July, generating the positive estimates for those two

[18] Given the recent literature on the bias of TWFE, I plan to test for conditional parallel trends with anticipation effects as well as the treatment effect of interest using the framework of Callaway and Sant'Anna (2021). Their approach would work well in this setting since they rely on never-treated units to estimate treatment effects. Therefore, I plan to utilize their R *did* package to estimate these effects in the future, better accounting for anticipation effects.

months. The negative estimates for August and September probably also come from the enforced delay in rice plantation that would have affected some farmers who would plant rice in early May otherwise. Finally, since there are very few fires to begin with in July, and since July happens to be the rainy season, the shifting of the wheat fire season perhaps does not have the same consequences for downwind pollution that the shifting of the rice season does.

I present robustness results to alternative specifications and sample selection in table 3.A1. These include the following: OLS estimation rather than Poisson, including fires data from 2000 and 2001,[19] and limiting the analysis to the sample for which GDP data is available.[20] The results are consistent with table 3.2 in all these robustness checks, with only the fire strength when limiting to the GDP sample becoming insignificant. This lack of power could be due to the effects of policy not having had enough time to accumulate by 2013 or to anticipation effects just before 2009. It should certainly not be taken as an indication that the laws did not increase November fire activity.

3.5.2 Effect of November fires on annual downwind pollution

Before turning to the causal effect of PM2.5 on district GDP, I discuss the effect of fire exposure on district PM2.5 levels. These results are equivalent to the first stage for the 2SLS results on GDP in the next section. As noted in the previous sections, fires in the winter are particularly harmful for PM2.5 levels due to prevailing meteorological conditions over North India that favor slower dispersion of the particulate matter over space. Further, the groundwater laws pushed agricultural fires in Punjab and Haryana toward November (early Winter). Therefore, I focus on the effect of November fire exposure on PM2.5.

The construction of the explanatory variable *log(Nov FRP exposure)* is described

[19] The NASA Aqua satellite was launched in 2002 and drastically improved estimates of fire activity in the FIRMS database

[20] 530 district between 2007-2013

in section 3.3.5. Referring to that section, variable F_{ot} is the total fire strength measured by Fire Radiative Power (FRP) of all fires in district o on day t in November. Certain fires can be stronger because more organic material is burnt, thereby producing higher amounts of particulate matter. Therefore I use FRP to maximize signal in the instrument relative to using count of fires.

Next, I implement various distance cut-offs on the exposure measure: origin districts at a larger distance than the cut-off are not used to construct FRP exposure for destination district. This is done for two reasons. Firstly, while wind fraction *times* inverse-distance weighting[21] captures some of the pollution decay over distance, it could miss out on some important features that govern decay, such as (i) rainfall along the path, which can cause the "wet deposition" of particulate matter (Vallero 2014) (ii) meteorological conditions along the path such as wind speed, temperature and relative humidity that could also alter the trajectory or cause further deposition out of the atmosphere and (iii) geographical features such as mountains along the way. For this reason, I hypothesize that larger cut-offs could add more noise to the instrument. Therefore, I test which distance cut-off maximizes the within-R2, in order to quantify the trade-off between signal and noise when increasing the distance cut-offs.

Table 3.3 shows results for cut-offs between 500 and 1000 km. In panel A, I present results from a fixed effects model that includes a district-specific time trend, equivalent to the first stage for equation 3.4. Panel B presents results from the first difference model with district fixed effects in equation 3.5, therefore assuming a district-specific trend in growth of PM2.5. Both these sets of results show strong and robust elasticities of PM2.5 to November FRP exposure, peaking at a cut-off of 900 km (for both the coefficient size and within-R2). The main result here is that a 1% increase in November FRP exposure increases PM2.5 levels by 0.029% (0.032%) with the FE (FD) model. It further illustrates the trade-off between sig-

[21] The distance decay could be modeled through a distance elasticity different from -1 too. I plan to do this later.

nal and noise when increasing distance to origin in constructing the instrument.[22] Globally, 900 km maximizes within-R2 when explaining PM2.5 using November FRP exposure. I therefore use that as the preferred distance to construct the instrument for PM2.5 in the next section.

Now, I address a concern that distance may be correlated with geographic determinants of PM2.5 (and GDP later). Controlling for district fixed effects in these regressions helps address that concern. But, distance also enters the instrument itself non-linearly; it may be that the district fixed effect does not fully address the issue. Therefore, I construct the instrument for each of these cut-offs by adding up wind fraction-weighted FRP from qualifying origins *without* inverse-distance weighting. Thus distance directly does not enter this instrument. Results for these regressions are presented in appendix table 3.A2. They do not suggest any cause for concern that distance entering the instrument non-linearly causes any bias in the first stage.

Lastly, in appendix table 3.A3, I confirm that higher FRP exposure only from fires during winter months affects annual PM2.5 levels. This can be explained by unfavorable meteorological conditions during winter that cause the particulate matter emissions from agricultural fires to stay suspended for longer. However, fires in the winter months other than November are not affected by the groundwater laws in Punjab and Haryana. Therefore, in order to quantify the effect of increased November fires due to the laws later, I use only November-based FRP exposure instrument in the analysis of the effect of PM2.5 on GDP in the next section.

3.5.3 Effect of PM2.5 on GDP

In this section, I turn to the impact of annual PM2.5 levels on annual GDP in Indian districts in panel A of table 3.4. I present results with the fixed effects in columns 1-3, and with the first difference specification in columns 4-5. Column 1 presents

[22] Results for regressions with a 100 km cut-off to no distance cut-off at all show an increasing within R2 until 900 km when they start dropping of monotonically.

the OLS estimate controlling for weather and including district and year fixed effects, but without district-specific linear time trends. The coefficient is positive and strongly significant. The causal effect of higher PM2.5 on GDP should be negative, given the harmful effects on human health and productivity, and potential effects on agriculture and machinery. The positive coefficient suggests that this specification is not sufficient to address the concern about omitted variables that jointly determine GDP and PM2.5, such as yearly demand shocks that cause higher GDP growth due to certain districts being more trade-exposed, for example. Higher economic activity in that year would increase PM2.5 levels in that district, and district fixed effects are insufficient to capture the co-movement of these variables. The estimate is biased upwards since the causal chain runs from GDP to PM2.5 in such cases. The sample period witnessed very strong GDP growth in Indian districts, making this a particular concern in this setting.

Column 2 presents results with the addition of a district-specific linear time trend to reduce this concern. The coefficient turns negative now, although it is imprecise, suggesting that this time trend is able to reduce the upward bias from the reverse causality of GDP to pollution. It also suggests the importance of including such time trends for non-stationary GDP data when focusing on the effect of jointly determined variables such as air pollution, as opposed to plausibly exogenous variables such as temperature deviations (Dell et al. 2012).

Before discussing the 2SLS estimates in columns 3 and 5, I focus on column 4 which presents the first difference estimate along with a district fixed effect, in effect assuming a district-specific trend in the growth rate of GDP. The coefficient is -0.03 and significant at the 5% level. The FD specification works much better with non-stationary data, and therefore this coefficient is less biased and also more precisely measured than the fixed effects regression with time trends in column 2.

Both these approaches solve some of the omitted variable problem plaguing estimation of the effect of PM2.5 on district GDP. However, any joint residual variation from the trend still causes upward bias in the estimates. I turn to the instrumental variable strategy to address this residual concern. In column 3, I present 2SLS results from the fixed effects model with district-specific time trends, instrumenting for PM2.5 using November FRP exposure with a 900 km distance cut-off. The estimate is now much larger, although the IV also increases standard errors as expected.

Panel B reproduces relevant first stage estimates from table 3.3. To test for weak instruments, I also present two statistics below the first stage estimates. Stock and Yogo (2005) suggest the use of the Cragg-Donald F-stat in a multivariate setting to test for weak instruments, with a rule of thumb that a value less than 10 indicates a potentially weak instrument. The Cragg-Donald F-stat is about 101.4; but this relies on iid assumptions for the errors. Therefore, I also report the Kleibergen-Paap (KP) F-stat which is equivalent to the robust F-stat with one endogenous regressor, as in this setting. The F-stat of 25.3 is comfortably above 10, and therefore concerns about weak instruments do not arise here.[23]

Column 5 presents the 2SLS results from the first difference model. The point estimate is slightly larger than column 3, and is estimated much more precisely. The KP F-stat is 26.4, again comfortably larger than 10. I consider the specification in column 5 as the preferred specification. These estimates suggest that increasing PM2.5 levels by 1% in a given year has a large negative causal effect of 0.18% on district GDP.

3.5.4 Quantifying the impact of groundwater laws on net GDP

The 2SLS estimates from the previous section can be used to estimate the effect of the increase in November fires in Punjab and Haryana on net GDP. Variation in

[23] Andrews et al. (2019) recommend the use of the effective F-statistic (MOP F-stat) of Olea and Pflueger (2013) in the case of a single endogenous regressor. This statistic is not easily calculated in any R or Stata package that implements IV with panel data. However, Andrews et al. (2019) also note that with one single endogenous regressor, the MOP F-stat is equivalent to the KP and robust F-stats. Therefore, the provided F-stat is the correct on to test for weak instruments. In future versions of the paper, also plan to present identification-robust Anderson-Rubin confidence intervals which are efficient regardless of the strength of the instrument.

the instrument in the previous section comes from November fires in all districts of India, both before and after passage of the groundwater laws. Therefore, we cannot interpret those estimates directly as the LATE associated with the laws. But we can use the estimates from this paper to calculate the percentage loss in net GDP across Indian districts in the following way.

Results from table 3.2 demonstrating that the laws increased November FRP in districts of Punjab and Haryana by an average of 0.542 log points. This increase in November FRP would increase FRP exposure within 900 km of each district. Since each Punjab and Haryana district sees the same proportionate increase in fires, and the inverse-distance and wind fraction weights do not change,[24] the proportionate increase in FRP exposure for all districts within 900 km is the same. Using the distance and wind fraction weights, this proportionate increase for each district within 900 km is also 0.542 log points. The increase in PM2.5 from this 0.542 log points higher November FRP exposure is 0.542*0.032 = 0.0173 log points, using the first stage estimate from column 5 of table 3.4. Finally, the proportionate reduction in GDP for each district is 0.0173*(-0.179) = -0.0031 log points or -0.3%.

The same proportionate reduction in GDP for districts within 900 km can produce different reduction in net GDP based on the initial GDP. This estimate for the average yearly impact of the groundwater laws on net GDP is -0.125%, 54.29 billion Indian Rupees or 1.12 billion USD (2004 values). This estimate is based on the 530 sample districts only, assuming that November fire exposure is limited to 900 km.

3.6 Conclusion

This paper estimates the unintended consequences of groundwater conservation laws in the two states of Punjab and Haryana on net Indian GDP, due to increased air pollution in downwind districts. In order to arrive at the net impact, I esti-

[24] I construct a 10-year average wind fraction for each day here

mate three elasticities. First, I provide evidence that the groundwater conservation laws shifted agricultural fire activity from late monsoon into early winter: biomass burnt in November increased by 72% while it decreased in October by 57%. This increased fire activity during winter more strongly affects PM2.5 levels because lower wind speeds and temperatures along with scant rainfall favor longer suspension of particulate matter in the smoke plumes.

Second, to quantify the impact of higher November fires on annual downwind PM2.5 levels, I construct a novel measure to summarize the exposure of each district to all upwind fires in November. I show that this exposure measure predicts 4% of the year-to-year variation in PM2.5 within each district. Third, I estimate the impact of higher PM2.5 levels on contemporaneous GDP. To solve concerns about omitted variable bias/reverse causality and non-stationarity, I adopt an identification strategy that relies on first differences to control for district-specific trends in PM and GDP, with an instrumental variable that provides exogenous variation in particulate matter. A 10% increase in PM2.5 levels in a given year reduces district GDP in that year by 1.8%. With these three elasticities, I calculate the yearly impact of increased PM2.5 levels due to the groundwater laws on net Indian GDP to be 0.125%. This estimate does not include non-monetized impacts of this pollution on health and well-being.

I have conducted some robustness checks that are presented in this chapter, and plan to conduct more. There are also some limitations to this approach. First, the estimate relies on the exposure instrument affecting downwind districts in accordance with its structure. While a chemical transport model could do better in modeling this relationship, it is much more resource-intensive to operate and may not do especially well for seasonal sources such as agricultural fires. Second, it is not a direct LATE of the legislation itself. It relies on the GDP elasticity of pollution that is estimated using November fires both within and outside Punjab and Haryana. In future work, I plan on directly estimating the impact of increased November fires from the groundwater laws on downwind GDP by restricting fires sources to Punjab and Haryana, and leaving out districts outside North India that the exposure instrument does not affect. While this would reduce power and potentially limit external validity to the rest of India, it will also allow me to estimate more directly the impact of groundwater laws on downwind GDP through exposure only to fires in Punjab and Haryana. I also plan to

On a different note, I also plan to explore the mechanisms behind the impact of the groundwater laws on net GDP, driven by the increase in November fires and PM2.5. Does this decrease come from a reduction in industrial production or agricultural output? Is the main channel the health and labor productivity impacts of PM2.5? Can firms adjust to this increased pollution by either moving or reallocating production to other months? One potential issue could be that legislation may affect local GDP in Punjab and Haryana through the costly adaptation to the laws themselves, biasing the estimates. Removing districts in these states from the sample would solve that problem. However, fire exposure is also likely to have the largest impact on PM2.5 on districts within these states and I prefer not to remove those districts from the sample for that reason. Instead, I intend to utilize outcomes such as the Index of Industrial Production that are not likely to be directly affected by the groundwater laws.

These laws were intended to conserve critical groundwater aquifers that have deplted at an alarming rate. In this paper, I do not investigate whether the laws increased groundwater levels, or quantify other benefits of this policy. I plan to do these in the future too. But the unintended consequences of this policy were to exacerbate the effects of agricultural fires on air pollution. Even though the use of fires to clear fields of residue has large costs in India, a combination of factors ranging from weak regulatory capacity or political capture by farm lobbies at the macro level, to credit constraints or lack of trust among smaller farmers may be responsible for this continued practice. By quantifying GDP losses in other states due to the increased pollution from fires in Punjab and Haryana, this paper suggests a mechanism whereby fiscal transfers from these downwind states affected by increased pollution could be made to farmers in Punjab and Haryana as payment not to burn (Jack et al. 2022).

The design of payments, specifically whether they should be upfront to alleviate credit constraints and combined with more stringent monitoring and enforcement, is another question for further research. While these payments go against the "polluter pays" principle that may be more relevant for the farmers in Punjab and Haryana, who are richer and have larger landholdings than the rest of India, any workable solutions in such second-best environments should consider the existing political and regulatory distortions which make these payments a sensible way to increase welfare. In the long term, incentivizing farmers to plant crops that are more suitable to the available resources, priced appropriately, could be a more sustainable solution.

3.7 Figures and Tables

Figure 3.1: Count of fires in Indian districts (2010) ↩

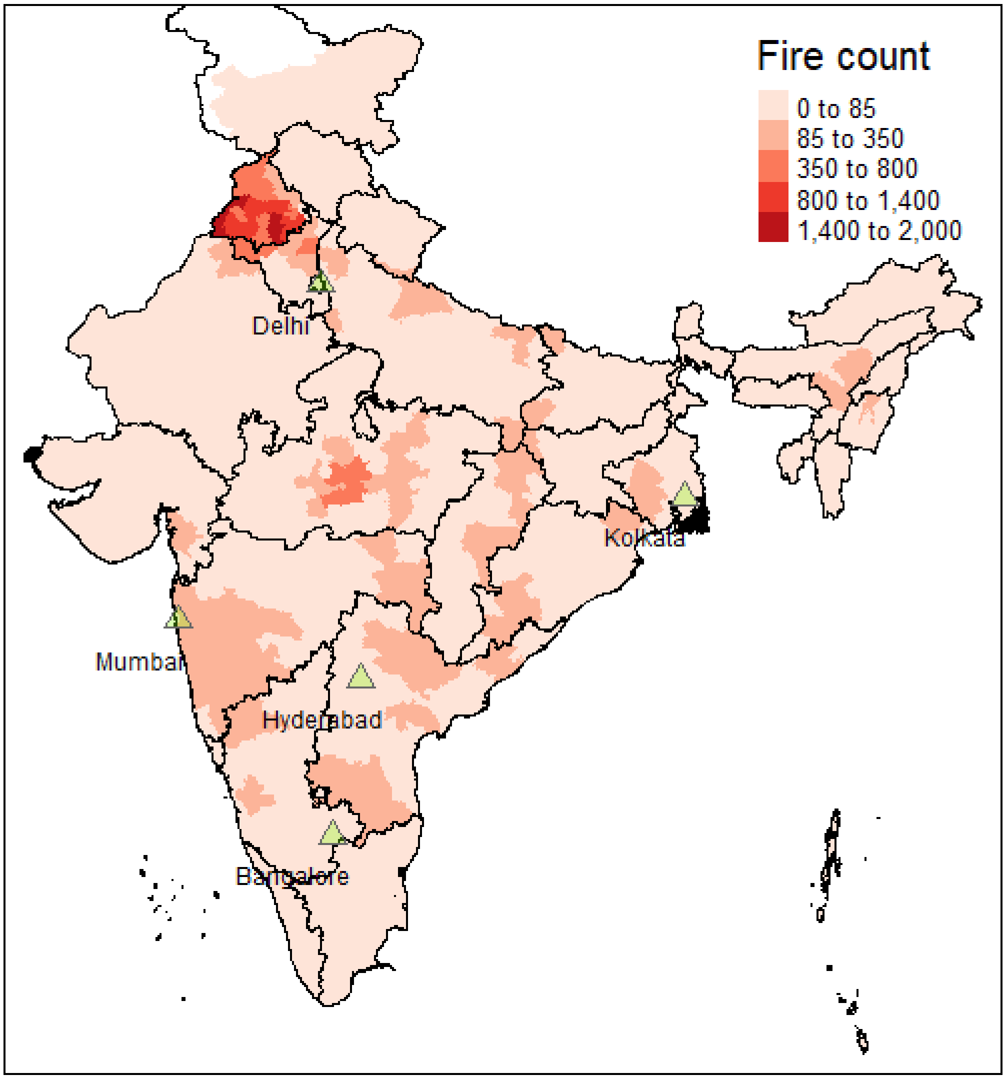
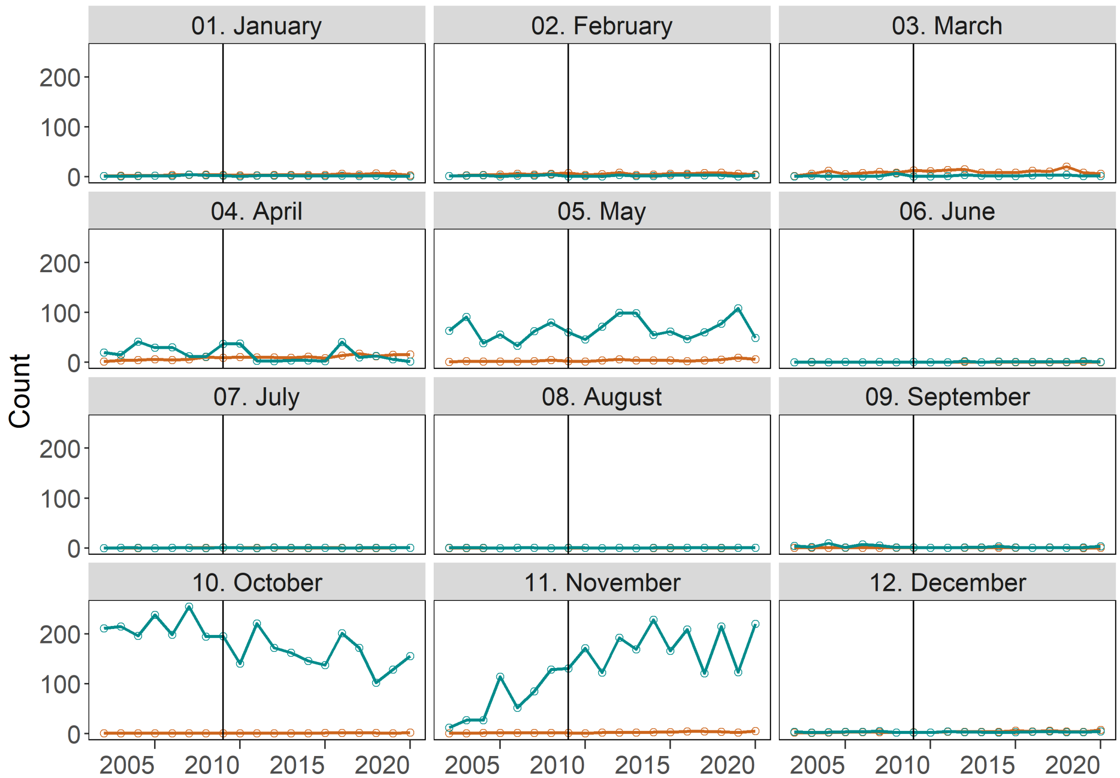


Figure 3.2: Trends in fire count and fire radiative power (2002-2020) ↩

(a) **Fire count**



(b) **Fire radiative power**

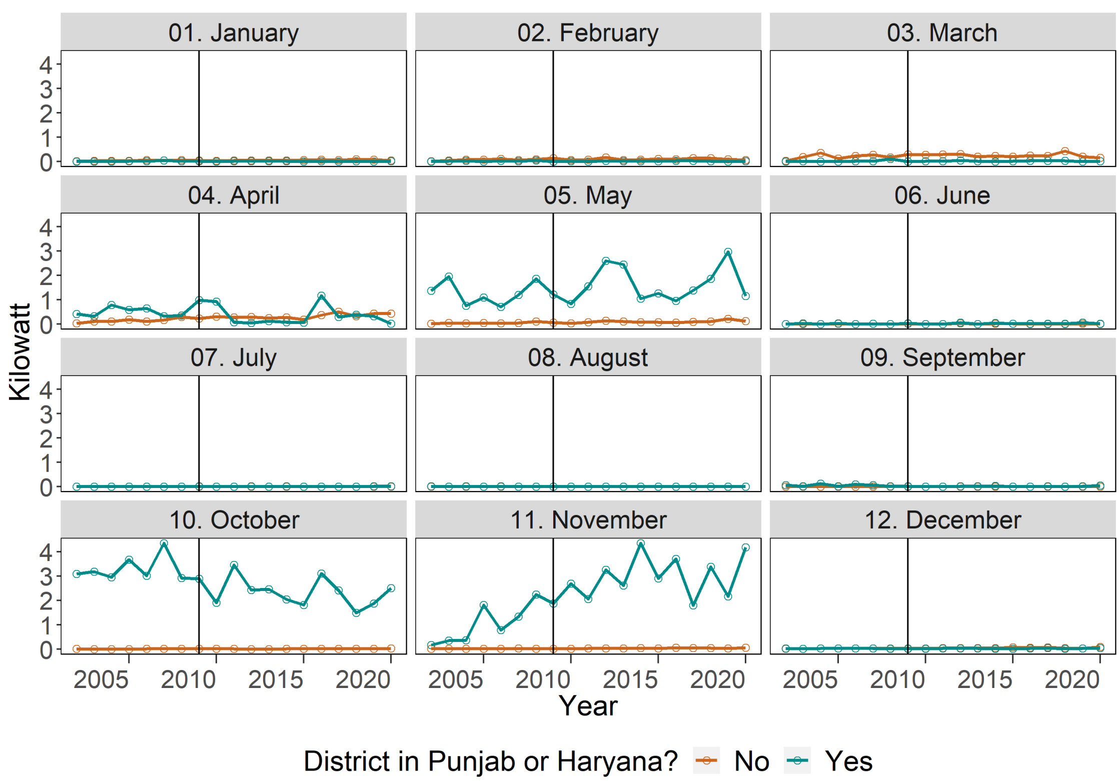
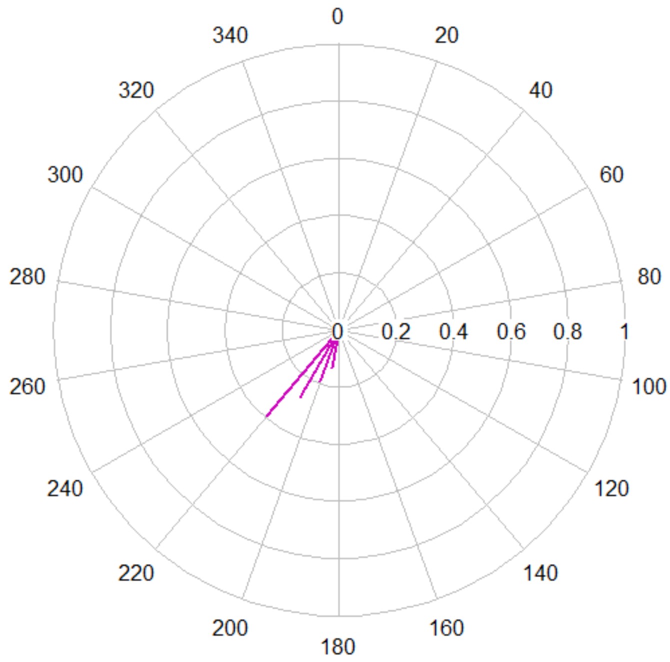
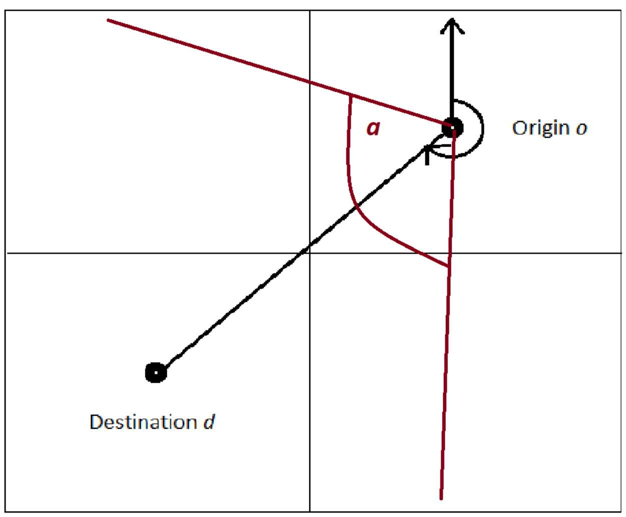


Figure 3.3: Construction of the fire exposure instrument ↩



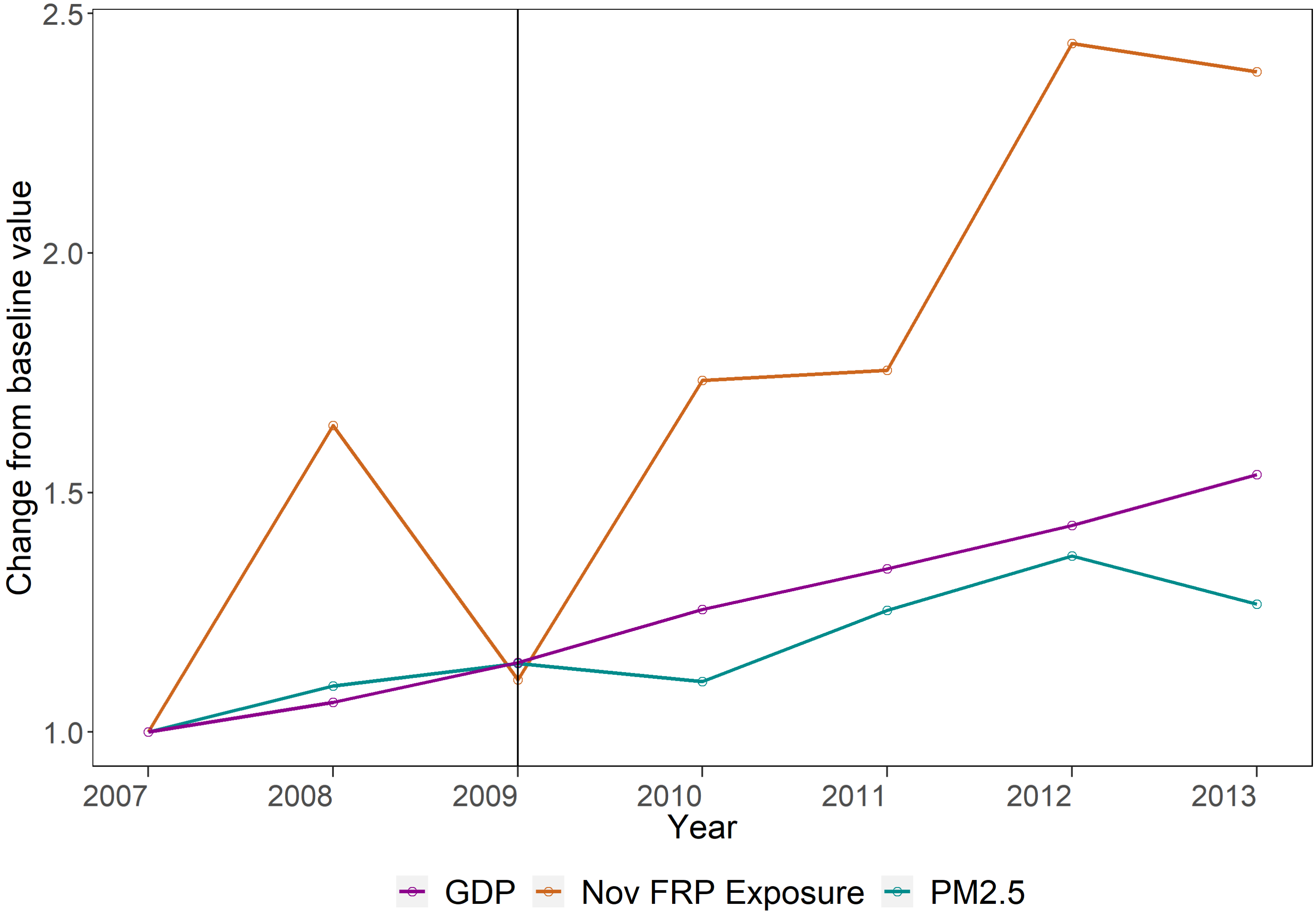
(a) Average wind directions at origin



(b) Direction from origin to destination

Note: The pink lines on top are fractions of time during the day when the wind was blowing in that bin.

Figure 3.4: Trend in fire exposure, PM and GDP (2007-2013) ↩



Note: Growth from the 2007 baseline value of each variable is plotted

Table 3.1: Summary Statistics

Variable	N	Mean	SD	Min	Max
Panel A: Monthly Fire Measures and Groundwater Law (2002-2020)					
Count of fires	143640	5.340	32.720	0	1148
Total Fire Radiative Power (mw)	143640	102.75	670.34	0	45044
Groundwater Law Dummy	143640	0.065	0.247	0	1
Panel B: Exposure to Upwind November Fires (FRP-based) with distance cut-off (2007-2013)					
Nov FRP exposure, cut-off = 500	3731	35.167	86.607	0.020	675.725
Nov FRP exposure, cut-off = 600	3731	38.924	87.728	0.053	675.902
Nov FRP exposure, cut-off = 700	3731	42.431	88.123	0.063	675.920
Nov FRP exposure, cut-off = 800	3731	45.904	88.126	0.085	675.985
Nov FRP exposure, cut-off = 900	3731	49.451	87.799	0.087	676.289
Nov FRP exposure, cut-off = 1000	3731	52.759	87.268	0.164	676.300
Panel C: Annual Particulate Matter and GDP (2007-2013)					
Mean PM2.5 (micrograms/m3)	3731	62.517	27.678	17.828	147.946
GDP (Billions of Rupees, Constant 2004)	3731	81.301	164.07	2.414	3728
Panel D: Annual Weather (2007-2013)					
Mean Temperature (°C)	3731	25.011	3.767	-10.369	29.847
Total Rainfall (mm)	3731	2165.9	430.47	0	2809
Mean Relative Humidity (Ratio)	3731	0.640	0.081	0.388	0.852
Mean Surface Pressure (kilo-pascal)	3731	96.85	4.946	56.460	100.83
Mean Windspeed (m/s)	3731	1.437	0.598	0.329	3.831

Notes: All data is at the district level. The sample consists of 530 districts, except for Panel A which consists of 630 districts (out of 640 census 2011 districts). The reduction is due to ICRISAT GDP data only being available between 2007-2013 for a subset of districts. ↩

Table 3.2: Poisson Estimates of Impact of Groundwater Laws on Monthly Fires

	Fire Count		Fire Radative Power	
	Pre-2009 Mean [SD]	(1)	Pre-2009 Mean [SD]	(2)
January	1.881	-0.749***	18.459	-0.746***
	[3.037]	(0.137)	[38.224]	(0.154)
February	2.384	-0.659**	25.36	-0.809***
	[3.759]	(0.278)	[55.396]	(0.218)
March	2.11	-0.529***	31.073	-0.771***
	[4.482]	(0.145)	[81.6]	(0.154)
April	20.527	-1.09***	440.866	-0.789***
	[27.907]	(0.260)	[601.855]	(0.266)
May	62.546	-0.430***	1330.916	-0.286*
	[72.912]	(0.118)	[1652.531]	(0.157)
June	0.494	0.253	13.74	0.040
	[1.306]	(0.181)	[55.281]	(0.158)
July	0.149	0.542***	2.401	0.726***
	[0.524]	(0.196)	[9.039]	(0.265)
August	0.36	-1.10***	6.031	-1.28***
	[1.077]	(0.289)	[19.759]	(0.249)
September	4.625	-1.83***	58.319	-1.94***
	[10.988]	(0.182)	[143.641]	(0.185)
October	192.287	-0.857***	2946.382	-0.855***
	[268.594]	(0.118)	[4473.063]	(0.158)
November	49.846	0.429***	788.759	0.542***
	[130.006]	(0.116)	[2265.141]	(0.159)

continued

December	3.084	-0.600***	26.838	-0.526***
	[3.997]	(0.162)	[40.39]	(0.175)
Observations	4018	140,372	4018	140,372
Pseudo R2		0.784		0.797
Years	2002-2018	2002-2018	2002-2018	2002-2018
Districts	41	630	41	630
State x Month FE		X		X
Year FE		X		X
District FE		X		X

Notes: Years 2002-2018. Columns 1 and 3 provide mean and SD of fire count and fire strength before 2009 **in Punjab and Haryana**. Columns labeled (1) and (2) provide Poisson estimates. Standard errors are clustered at district and State x Year. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 3.3: Impact of distance-weighted November fire exposure on PM2.5

	Dependent Variable: log(PM)					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Fixed Effects Model						
log(Nov FRP Exposure)	0.011***	0.016***	0.022***	0.027***	0.029***	0.028***
	(0.004)	(0.004)	(0.005)	(0.006)	(0.006)	(0.006)
Observations	3,731	3,731	3,731	3,731	3,731	3,731
Within R2	0.539	0.542	0.546	0.550	0.551	0.549
Panel B: First Differences Model						
log(Nov FRP Exposure)	0.008***	0.009***	0.010***	0.031***	0.032***	0.031***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)
Observations	3,178	3,178	3,178	3,201	3,201	3,201
Within R2	0.171	0.172	0.173	0.197	0.197	0.193
Distance Cutoff	[500 km]	[600 km]	[700 km]	[800 km]	[900 km]	[1000 km]
Weather Controls	X	X	X	X	X	X
District and Year FE	X	X	X	X	X	X
District x Time Trend	X	X	X	X	X	X

Notes: Years 2007-2013. The sample is limited to districts for which GDP data is available. Each column of Panel A and B provides estimates from the same regression specification but with a different distance cut-off when constructing the FRP exposure instrument. Estimates in each panel are equivalent to the first stage for columns 3 and 5 in table 3.4. Standard errors are clustered at district and Region x Year. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 3.4: Impact of Air Pollution (PM2.5) on GDP

	Dependent Variable				
	log(GDP)			Δ log(GDP)	
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS and 2SLS Results					
log(PM2.5)	0.147^{***}	-0.008	-0.159	-0.030^{**}	-0.179^{***}
	(0.035)	(0.016)	(0.097)	(0.014)	(0.069)
Observations	3,731	3,731	3,731	3,201	3,201
R2	0.996	0.999	0.999	0.379	0.326
Weather Controls	X	X	X	X	X
District and Year FE	X	X	X	X	X
District x Time Trend		X	X		
First Differences				X	X
2SLS Estimate			X		X
Panel B: First Stage Results					
log(Nov FRP Exposure)			0.029^{***}		0.032^{***}
			(0.006)		(0.006)
Cragg-Donald F-stat			101.4		116.5
Kleibergen-Paap F-stat			25.3		26.4

Notes: Years 2007-2013. The sample is limited to districts for which GDP data is available. Panel A, columns 1-3, show estimates for both OLS and 2SLS regressions of log GDP level on log PM2.5, starting without a time trend, then controlling for a time trend and finally conducting 2SLS with time trend. Columns 4 of panel A shows an OLS estimate using first differences while column 5 instruments for first difference of log PM with first difference of log Nov Exposure (900 km cut-off). Standard errors are clustered at district and Region x Year. *p<0.1; **p<0.05; ***p<0.01. ↩

3.8 Appendix

Table 3.A1: Impact of Groundwater Laws on Monthly Fires in Punjab and Haryana - Robustness

	Fire Count			Fire Radative Power		
	(1)	(2)	(3)	(4)	(5)	(6)
January	-0.143**	-0.747***	-0.746***	-0.164***	-0.765***	-0.918***
	(0.059)	(0.133)	(0.078)	(0.054)	(0.156)	(0.080)
February	-0.368***	-0.737***	-0.786***	-0.407***	-0.867***	-0.949***
	(0.080)	(0.252)	(0.205)	(0.077)	(0.202)	(0.229)
March	-0.398***	-0.668***	-0.914***	-0.500***	-0.957***	-1.27***
	(0.049)	(0.142)	(0.080)	(0.057)	(0.156)	(0.137)
April	-0.790***	-1.06***	0.198	-0.735**	-0.766***	0.028
	(0.289)	(0.254)	(0.186)	(0.335)	(0.263)	(0.183)
May	-0.118*	-0.534***	-0.097	-0.072	-0.410***	-0.086
	(0.059)	(0.118)	(0.076)	(0.072)	(0.158)	(0.080)
June	-0.078	0.243	0.591***	-0.233***	0.054	0.577***
	(0.052)	(0.165)	(0.170)	(0.077)	(0.157)	(0.118)
July	-0.152*	0.517***	1.20***	-0.159***	0.698***	1.55***
	(0.076)	(0.150)	(0.401)	(0.055)	(0.220)	(0.224)
August	-0.495***	-1.05***	-0.035	-0.699***	-1.25***	0.333*
	(0.082)	(0.295)	(0.070)	(0.065)	(0.254)	(0.182)
September	-1.09***	-1.95***	-1.42***	-1.28***	-2.07***	-1.52***
	(0.258)	(0.193)	(0.144)	(0.250)	(0.195)	(0.141)
October	-0.223***	-0.817***	-0.380***	-0.190*	-0.821***	-0.535***
	(0.074)	(0.119)	(0.070)	(0.107)	(0.159)	(0.079)
November	1.05***	0.509***	0.238***	1.15**	0.613***	0.124

continued

	(0.382)	(0.120)	(0.081)	(0.430)	(0.161)	(0.091)
December	-0.370**	-0.677***	-0.323**	-0.359*	-0.640***	-0.355***
	(0.154)	(0.157)	(0.132)	(0.178)	(0.166)	(0.103)
Observations	56,082	149,257	43,904	56,082	149,257	43,904
Specification	OLS	Poisson	Poisson	OLS	Poisson	Poisson
Years	2002-2018	2000-2018	2007-2013	2002-2018	2000-2018	2007-2013
Districts	630	630	630	630	630	630
State x Month FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
District FE	X	X	X	X	X	X

Notes: Provides robustness checks for table 3.2. Columns 1 and 4 conduct OLS estimation with log(fire count) and log(FRP) as the dependent variables. Columns 2 and 5 conduct the Poisson estimation with fires data from 2000 and 2001, when the fires are less reliable. Columns 3 and 6 conduct Poisson estimation by restricting sample to data from the 530 districts over 2007-2013 which have GDP data available. Standard errors are clustered at district and State x Year. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 3.A2: Impact of November fire exposure without distance weighting on PM2.5

	Dependent Variable: log(PM)					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Fixed Effects Model						
log(Nov FRP Exposure)	0.014***	0.021***	0.028***	0.032***	0.032***	0.028***
	(0.004)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Observations	3,731	3,731	3,731	3,731	3,731	3,731
Within R2	0.542	0.545	0.551	0.554	0.553	0.549
Panel B: First Differences Model						
log(Nov FRP Exposure)	0.010***	0.011***	0.011***	0.034***	0.033***	0.030***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)
Observations	3,178	3,178	3,178	3,201	3,201	3,201
Within R2	0.173	0.174	0.174	0.201	0.198	0.191
Distance Cutoff	[500 km]	[600 km]	[700 km]	[800 km]	[900 km]	[1000 km]
Weather Controls	X	X	X	X	X	X
District and Year FE	X	X	X	X	X	X
District x Time Trend	X	X	X	X	X	X

Notes: Years 2007-2013. Robustness to dropping distance from construction of exposure instrument in table 3.3. The sample is limited to districts for which GDP data is available. Each column of Panel A and B provides estimates from the same regression specification but with a different distance cut-off when constructing the FRP exposure instrument. Standard errors are clustered at district and Region x Year. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 3.A3: Impact of distance-weighted Monthly fire exposure on annual PM2.5

	Dependent Variable: log(PM)					
Exposure Month	Jan	Feb	Mar	Apr	May	Jun
Panel A: Estimates for January to June						
log(Monthly FRP Exposure)	0.007	0.017***	0.011**	-0.002	0.007	-0.006***
	(0.006)	(0.006)	(0.004)	(0.006)	(0.006)	(0.002)
Observations	3,731	3,731	3,731	3,731	3,731	3,718
Within R2	0.535	0.540	0.537	0.534	0.535	0.537
	Dependent Variable: log(PM)					
Exposure Month	Jul	Aug	Sep	Oct	Nov	Dec
Panel B: Estimates for July to December						
log(Monthly FRP Exposure)	-0.004	-0.002	0.003	-0.010*	0.029***	0.012*
	(0.003)	(0.003)	(0.003)	(0.005)	(0.006)	(0.007)
Observations	3,697	3,726	3,730	3,731	3,731	3,731
Within R2	0.534	0.534	0.535	0.536	0.551	0.536
Distance Cutoff	[900 km]	[900 km]	[900]	[900 km]	[900 km]	[900 km]
Weather Controls	X	X	X	X	X	X
District and Year FE	X	X	X	X	X	X
District x Time Trend	X	X	X	X	X	X

Notes: Years 2007-2013. The sample is limited to districts for which GDP data is available. Each single column in Panel A and B displays estimates for the regression of annual PM2.5 on exposure to fires during that month of the year only, using the same specification as in table 3.3. Standard errors are clustered at district and Region x Year. *p<0.1; **p<0.05; ***p<0.01. ↩

Chapter 4

Industrial Water Pollution and Agricultural Production

4.1 Introduction

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

Even in high-income countries, the social costs of water pollution have been challenging to quantify. While surveys show high levels of public interest in water quality, research has rarely found economically significant impacts of water pollution. This could be because the costs truly are low, or alternatively because water pollution is especially difficult to study. Low quality and availability of pollution measurements, the difficulty of modeling complex spatial relationships, and the wide variety of distinct pollutants may have both inhibited research and attenuated estimates that do exist (Keiser and Shapiro 2017).

This paper estimates the effects of industrial water pollution on agricultural production in India. We study agriculture because several reasons suggest it could be the site of large aggregate effects of water pollution. Agriculture uses four times more water than all other sectors of the economy combined (FAO 2018), and irrigation water is rarely treated before use, unlike drinking water. The agricultural sector is also large and ubiquitous, so it can be found near virtually every source of pollution. We focus on 71 industrial sites identified by India's Central Pollution Control Board in 2009 as "severely polluted" with respect to water pollution, out of 88 sites selected for intensive study. India's industrial clusters are home to some of the greatest concentrations of industrial pollution in the world (Mohan 2021), so if industrial water pollution matters anywhere, it likely matters here.

Our research design exploits the fact that water pollution, unlike air pollution, almost always flows in only one direction from its source. When industrial wastewater is released into a flowing river, it creates a spatial discontinuity in pollution concentrations along that river. Areas immediately downstream of a heavily polluting industrial site will have higher pollution levels than areas immediately upstream, yet they are likely similar otherwise. This makes upstream areas a reasonable counterfactual for the downstream areas in studying the impacts of water pollution on economic outcomes.

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

To measure agricultural outcomes, our main analysis relies on satellite data. No other data source is available at high enough spatial resolution to enable a spatial regression discontinuity design; even in the United States, aggregate statistics are too coarse and agricultural surveys too sparse. We use basic hydrological modeling to define precise upstream/downstream relationships between industrial sites and millions of pixels of satellite imagery. As proxies for agricultural output, we use remote sensing products developed by earth scientists to measure vegetation density, plant health, and metabolic activity. The measure we focus on is the normalized differenced vegetation index (NDVI); we also examine two other measures but find them to be noisier. All three have been shown to reliably predict crop yields across a range of settings (Running et al. 2004; Burke and Lobell 2017; Lobell et al. 2020, 2022; Asher and Novosad 2020). We also show in our context that NDVI predicts agricultural output in aggregate statistics.

We show three sets of results. First, we quantify the water pollution released by India's "severely polluted" industrial sites, using the available monitoring station data. We show that there is a large, discontinuous increase in water pollution at these exact locations, raising prior levels of pollution in nearby rivers by 140%. Second, we find that remote sensing measures of crop growth are lower downstream of high-polluting industrial sites, but only by 2.6 percent. The estimates are precise; confidence intervals exclude differences of more than 5 percent. A rough conversion implies that the sites reduce true crop yields by about 1 percent, suggesting that even the localized effects of industrial water pollution are small. Third, we document that farmers are neither substituting factors of production away from agriculture nor applying additional compensating inputs. The effects of being downstream on crop area, irrigation, labor, and population are small and statistically insignificant.

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

This paper contributes to the existing literature in four ways. First, it provides among the first quasi-experimental evidence on any form of costs of industrial water pollution. In recent papers on India, Do et al. (2018) study the effects of industrial water pollution on infant mortality, while Brainerd and Menon (2014) study the effects of agricultural water pollution to child health. In the United States, Keiser and Shapiro (2017) study the effect of all water pollution on property values. Most other economics literature on the costs of water pollution deals with domestic water pollution in the context of providing clean drinking water. Second, this paper documents economic costs of pollution to agriculture; to our knowledge only one other paper does so quasi-experimentally, but in the context of air pollution Aragón and Rud (2016). Third, it adds to the small but rapidly growing literature on the costs of pollution in developing countries (Jayachandran 2009; Y. Chen et al. 2013; Adhvaryu et al. 2022).

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

4.2 Background

4.2.1 Industrial water pollution and crop growth

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

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

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

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

4.2.2 Remote sensing of crop yields

In order to quantify the effect of industrial water pollution on agricultural output, the ideal data would be at a spatial resolution of tens of meters, similar to the Cropland data layer from the US Department of Agriculture. However, the most granular spatial extent over which Indian agricultural data is collected and reported is the administrative unit of a district, an entity that is about 100 sq km on average. This is far too large for our purposes since water pollution may get diluted over distance; also, within-district sources of pollution may affect only part of the district whereas the agricultural output data are for the district as a whole. We solve this challenge by utilizing agricultural proxies from remote sensing data that have been widely used in the scientific agronomic literature to measure crop yields (Running et al. 2004; Burke and Lobell 2017; Lobell et al. 2022), and are increasingly common within the economics literature as well (Asher and Novosad 2020; Lobell et al. 2020).

The two most commonly used and related vegetation indices (VIs) to proxy for agricultural yields are the Normalized Difference in Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI). These indices aim to capture the amount of photosynthetic activity in plants in the following way. The chlorophyll pigment that gives leaves their green color absorbs much of the red light in the visible spectrum in healthy plants. Other cell structures of the same plant reflect most of the near-infrared light in the invisible part of the electromagnetic spectrum. A healthy plant with high photosynthetic activity due to high amounts of the Chlorophyll pigment will reflect less red light and more near-infrared light. The NDVI provides a simple mathematical formula to compare these two reflectance values and thereby establish the strength of photosynthetic activity in plant matter, independent of other land cover. EVI is very similar but uses additional information from the blue part of the electromagnetic spectrum to reduce atmospheric interference and the influence of background vegetation (Son et al. 2014). NDVI is strictly bound by the interval [-1, 1] whereas the EVI may have values outside that range, although in practice this is rare.

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

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

4.3 Research Design

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

Figure 1 illustrates this sharp discontinuity. It is an aerial photograph of one site in our sample: the Nazafgarh Drain Basin on the Yamuna River just north of New Delhi. The river flows from north to south and enters the image at the top with a green color. In the center of the image, an industrial effluent channel meets the river, discontinuously turning the river black. Although color is neither a sufficient nor necessary condition for any specific pollutant, the color difference confirms the presence of water from a different source, and color is correlated with water pollution. Remote sensing measures, which include visible light as well as a broader range of wavelengths, are becoming increasingly common in water quality monitoring (Gholizadeh et al. 2016).

4.3.1 Sample selection and treatment definition

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

The first challenge is that to define “downstream” villages, we need to introduce a channel of exposure through which river pollution reaches the villages. Geometrically, the set of points downstream of a point source is a line – the path water follows to reach the ocean – not an area. Possible exposure channels are through (a) surface water irrigation, using water pumped directly from a river; (b) surface water irrigation, using water from a canal that diverts water from the river; (c) groundwater irrigation, using water pumped from underground aquifers that may have been contaminated either through direct seepage or from surface water sources; or (d) soil contamination, from groundwater in areas with high water tables. Each of these exposure channels produces unique spatial patterns of treatment intensity, depending on topography, geology, soils, infrastructure, and irrigation practices.

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

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

Our solution to these challenges is illustrated in Figure 3. This figure shows our research design for one site in our sample: Bhillai-Durg, a major industrial city in the state of Chhattisgarh. The center of this industrial site is represented by the orange dot. First, we follow the site downstream a short distance (25 km, to the upper yellow dot). Second, we follow this point upstream into the uppermost reaches of its watershed (to the lower yellow dot). Third, we find the downstream flow path of this "headwater" point. This headwater flow path forms the base of our analysis. We define our sample as all villages within 25 km of this flow path. We define treatment status by projecting (snapping) villages and the industrial site onto the headwater flow path, and calculating the flow distance between them along this path. Villages are assigned to downstream if their projection is downstream of the projected industrial site and upstream otherwise.

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

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

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

4.3.2 Regression discontinuity

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

$$y_{ist} = \beta Downstream_{is} + \gamma Distance_{is} + \delta Distance_{is} \times Downstream_{is} + \alpha_{st} + \varepsilon_{ist} \tag{4.1}$$

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

The coefficient of interest is β, the effect of being downstream of an industrial site. The running variable is downstream distance along the river flow path, defined such that each industrial site is at zero. Positive values indicate that a village is downstream of the industrial site; negative values indicate that the village is upstream. We include site-by-year fixed effects α_{st} so that the treatment effect at the discontinuity is identified only using variation between upstream and downstream observations for the same industrial site in the same year. For pollution outcomes, all details are identical, except that i represents a water quality monitoring station instead of a village.

We estimate local linear regressions on each side of the cutoff without higher order polynomials, following Gelman and Imbens (2014). We report results using a range of bandwidths with a minimum value of 25 km. Smaller bandwidths might fail to include villages fully exposed to pollution, due to the way we construct our sample. In future work we will implement the optimal bandwidth algorithm of Calonico et al. (2020). We use a triangular kernel, which is optimal for estimating local linear regressions at a boundary (Fan and Gijbels 1996). We cluster standard errors by district (the administrative level below state) to account for correlation across space and time. Clustering also accounts for repeated observations, when the same village appears more than once in the stacked sample for different industrial sites. Finally, we weight village observations by crop area so that our results represent the treatment effects for the average acre of cropland, which is more easily interpretable than effects for the average village.

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

The identifying assumption for this RD design is that the upstream patterns in pollution and agricultural outcomes would have continued smoothly downstream if the industrial site did not exist. Our samples represent continuous swaths of land area, making it a priori unlikely that there would be discontinuities in either river pollution or agricultural outcomes. One way the assumption would be violated is if industrial sites were strategically located downstream of the best agricultural land. Most of the sites in our sample are part of cities and towns that arose through usual agglomeration processes, and we can test for discontinuities in land quality.

Another way the assumption would be violated is if there is sorting of agricultural inputs or farmers themselves. Migration and/or disinvestment in downstream areas is possible, and we can test for it. These resources are unlikely to shift to the areas immediately upstream, rather than urban areas elsewhere, given India's rigid land and labor markets (Hsieh and Klenow 2009; Duranton et al. 2016).

4.3.3 Limitations of temporal variation

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

One, the starkest variation in our context is spatial, not temporal. Our causal identification is based on the location of industrial sites, which are extremely persistent and have not changed for decades. Most of these sites have grown over time, but this growth is correlated across sites over time as India has industrialized, leaving little useful variation. Two, available measures of industrial plant growth are noisy. The Economic Census gives the number of, and employment in, high-polluting plants in a town or village, but is known to suffer from data quality limitations (Bardhan 2013). Three, pollution itself cannot be used as an independent variable without an instrument. Pollution concentrations are strongly affected by changes in runoff (as varying volumes of water dilute the same pollution *load*), which itself strongly affects agricultural production through availability of irrigation water.

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

4.4 Data

4.4.1 Sources

4.4.1.1 Industrial sites

The Central Pollution Control Board (CPCB) selected 88 industrial sites for detailed, long-term study in 2009. Names of these sites were taken from the CPCB document (Central Pollution Control Board 2009a). We identified the geolocation of each site using Google Earth and other publicly available reference information. These sites are displayed as orange dots in Figure 2.

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

4.4.1.2 Pollution measurements

We use a rich dataset of water pollution measurements along rivers in India collected by the Central Pollution Control Board, as collected and published by Greenstone and Hanna (2014). This dataset includes monthly observations from 459 monitoring stations along 145 rivers between the years 1986 and 2005. We extend this data set by downloading yearly pollution readings for the same stations from 2006-2012 from the Cental Pollution Control Board's website. Then we construct yearly averages for the pre-2005 data and append these to the newly downloaded data.

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

Many water quality parameters have been collected by the CPCB at some point. However, only a few parameters are measured consistently. We focus on chemical oxygen demand (COD), a standardized laboratory test that serves as an omnibus measure of organic compounds, which industrial plants typically generate in high quantities. Along with the related but narrower test of biochemical oxygen demand (BOD), COD is the Indian government's top priority in regulating industrial wastewater (Duflo et al. 2013). We also report three other widely reported measures: BOD, dissolved oxygen (DO), and electrical conductivity (EC), a measure of salinity.

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

4.4.1.3 Agricultural outcomes via remote sensing

As discussed in the background section, we utilize remote sensing data to construct measures of agricultural outcomes. All of our remote sensing data retrieval and processing are carried out within the Google Earth Engine. For the NDVI and EVI calculations, we utilize data from the Landsat 7 satellite launched by NASA and operated by the US Geological Survey (USGS). The temporal coverage of Landsat 7 from 1999 onward suits our analysis better than the newer Landsat 8 that was launched in 2013. Landsat 7 covers each point on earth every 16 days, thereby providing an image of the globe at that frequency. The spatial resolution of 30m of Landsat 7 is also superior to the 250m resolution of the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard NASA's Terra satellite, data from which are also commonly used to calculate these vegetation indices. We utilize the Surface Reflectance product that is recommended for constructing vegetation indices since it includes atmospheric corrections for aerosols, and apply the quality assurance mask that indicates cloud cover over the pixel (Young et al. 2017). We also apply an agricultural land use mask from the Copernicus Global Land Service (CGLS) to ensure that only pixels where agricultural activity is being carried out are included in the sample.

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

We spatially aggregate these pixel-level data to the village-level to match with

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

4.4.1.4 Agricultural outcomes in aggregate data

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

Our primary outcome of interest is agricultural revenue. To calculate this, we multiply the quantity of each of 16 crops available in the dataset by the mean price for that crop in that district between 1990-2002. For districts without price data, we impute the state mean if available or the national mean otherwise.

[1] Available for download from the SHRUG platform.
[2] http://data.icrisat.org/dld/src/crops.html

4.4.1.5 Village shapefiles and covariates

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

Boundaries of villages and towns may change over time. Here we rely on the Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG) provided by the Development Data Lab.[4] The SHRUG provides an identifier called a 'shrid' for a group of contiguous villages or towns that can be combined into unchanged spatial entities over three population censuses (1991, 2001, 2011). Village/town level administrative data from various censuses and surveys can be linked to each shrid, after aggregating over the appropriate spatial extent. Almost 96% of villages from the 2001 population match a single shrid, therefore not needing any spatial aggregation over village polygons or administrative data. For the rest of the villages, we dissolve the polygons boundaries to obtain the shrid boundaries, and aggregate administrative data over the villages within each shrid.

Some baseline village covariates and outcomes are summarized in table 4.8. We use more covariates to test for continuity at the RD threshold later on. These include, but are not limited to, total population and area of the village; irrigated area (gross irrigated area per gross cropped area, percent); river, canal and groundwater specific measures of the previous variable; and land cropped (net cropped area per

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

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

total district area, percent).

4.4.2 Hydrological modeling

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

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

Headwater points. To generate a "headwater" point for each industrial site, we use the Trace Downstream tool in ArcGIS Pro (from the Ready to Use Hydrology toolset) to find the flow path of each industrial site. This flow path is the route that effluent released at the site must follow to reach the ocean. We then find the point on this flow path that is 25 km downstream of the site. Next, we use the Watershed tool (in the same ArcGIS Pro toolset) to find the area that drains into that point. We find the flow lengths of all villages within this watershed by intersecting the watershed polygon with village centroids and matching village centroids to the flow length raster. We identify the longest possible flow path within this watershed by choosing the village at the 95th percentile of flow length within this set. We choose the 95th percentile instead of the maximum to avoid erroneous values that sometimes arise at the edges of the watershed polygons.

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

Village distance and treatment status. To calculate distances for the RD design, we project all village centroids, industrial sites, and monitoring stations into one-dimensional river space. Specifically, we snap all these points to the nearest point along the headwater flow path. We then find the flow length (i.e., to the ocean) of each snapped point by matching it to the flow length raster. We construct distance, the running variable, as the difference in flow lengths between each village or monitoring station and the corresponding industrial site. We also construct a downstream indicator variable equaling one if the distance variable is positive, meaning that the village or station is downstream of the industrial site.

4.4.3 Continuity tests and summary statistics

We provide summary statistics in Table 4.8 for our main outcome variables on pollution and agricultural output, and in the first column of Appendix Table 6 for covariates.

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

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

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

Standard errors and RD point estimates for these covariates and many others are shown in Appendix Table 6 using a range of bandwidths. Across the 31 variables and 3 bandwidths we test, few coefficients are statistically significant. Taken together, there is little evidence to suggest that agricultural outcomes would be different immediately downstream of the industrial sites if they did not exist. It also does not appear that commercial and public amenities or demographic characteristics of villages are affected by being downstream of these industrial sites. In robustness checks, we control for all these covariates.

4.5 Results

4.5.1 Pollution

We first show that the industrial sites considered “severely polluted” by the Central Pollution Control Board do in fact increase pollution levels discontinuously in nearby rivers. The amount of water pollution released by these sites has not previously been quantified in publicly available sources.

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

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

Table 2 quantifies these results. It reports RD estimates from separate regressions for each parameter, for bandwidths of 25, 50, and 100 km. Dependent variables are listed in rows; each cell shows the estimated coefficient on the Downstream indicator variable, controlling for distance on each side of the industrial site along with site-by-year fixed effects.

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

4.5.2 Agricultural outcomes

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

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

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

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

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

4.5.3 Agricultural inputs and economic outcomes

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

Table 4, Panel B reports RD estimates for a set of agricultural inputs. Labor, as measured by the share of employment in agriculture, does not change immediately downstream of severely-polluting industrial sites (the point estimate is small and not statistically significant). Neither does land, as measured by crop area under cultivation. Irrigation plausibly might either increase to compensate for damage from pollution or decrease because the water itself is harmful, but that does not appear to happen overall (share of crop area under irrigation) nor for any particular source (share of irrigation from rivers, canals, or wells).

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

4.6 Explanations

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

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

Industrial pollution has beneficial components that balance the harms. Industrial effluent often includes salinity, heavy metals, and other components that are known to harm crops. However, they can also include nitrates, phosphates, and potassium, which can benefit plants as nutrients. It is possible that the net effects of industrial effluent are near zero, even if individual components have positive and negative effects. It is also possible that our estimates, which average over across industrial sites, mask heterogeneity across sites. In future work, we will estimate heterogeneous treatment effects by site and investigate whether the limited pollution data can shed any light on the differences in pollutants across sites.

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

Pollution harms output quality, not quantity. It is possible that industrial water pollution does harm crops, but only in ways that affect crop quality rather than quantity. For example, a crop such as rice might absorb heavy metals, bringing adverse health effects to consumers but leaving yield unaffected. Obvious quality effects may capitalize into prices; other quality effects may not. In future work, we will test for impacts to crop prices in aggregate data, as well as local human health.

Remote sensing proxies are unable to detect yield effects from water pollution. Despite a range of papers in both the economics and scientific literatures that have found satellite-derived vegetation indicies to be useful proxies for crop yields and agricultural output, many questions and uncertainties remain about their capabilities. One possibility is that NDVI and EVI are simply not well-suited to pick up the effects of industrial water pollution on crops. This could be because water pollution affects crops in ways that do not show up in remote sensing measures, although many of the agronomy studies on water pollution do specifically report negative impacts to leaf size and color, characteristics that vegetation indices are well-tailored to measure. It may also be possible that farmers adjust crop choice in response to pollution exposure, even though we do not see other inputs change. NDVI and EVI are affected by vegetation type in addition to crop health, so if farmers switch to new crops with greater baseline biomass or leaf canopy, it could offset the direct harms from pollution.[5]

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

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

4.7 Conclusion

This paper provides the first quasi-experimental evidence on the costs of industrial water pollution to agriculture. We examine 71 industrial sites in India identified by the government as "severely polluting" and estimate the costs of their pollution to downstream agriculture. Our regression discontinuity research design exploits the unidirectional flow of water pollution along with the location of these severely polluted industrial sites. To overcome the limitations placed by spatially aggregated administrative data on agricultural output, we construct remote sensing proxies for agricultural yields including Normalized Difference Vegetation Index, Enhanced Vegetation Index and Net Primary Production. These proxies have been shows to perform well in predicting yields both in the scientific and economics literature, and we verify that they predict agricultural yields within our sample too. We also extend the data set on water pollution collected by Greenstone and Hanna (2014) between 1986-2005 to 2012 using yearly data available from the Central Pollution Control Board of India.

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

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

4.8 Figures and Tables

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

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

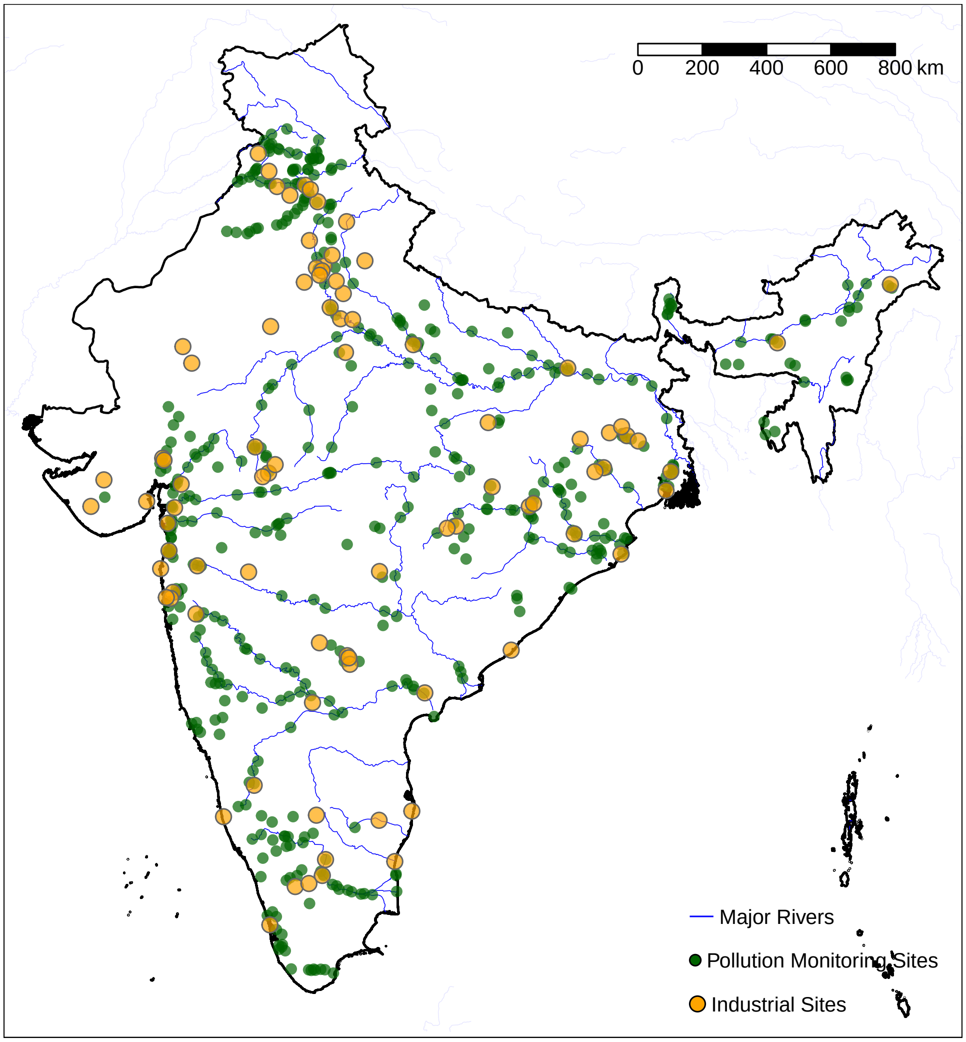


Figure 3: Illustration of the sample selection and treatment assignment for our main research design. ↩

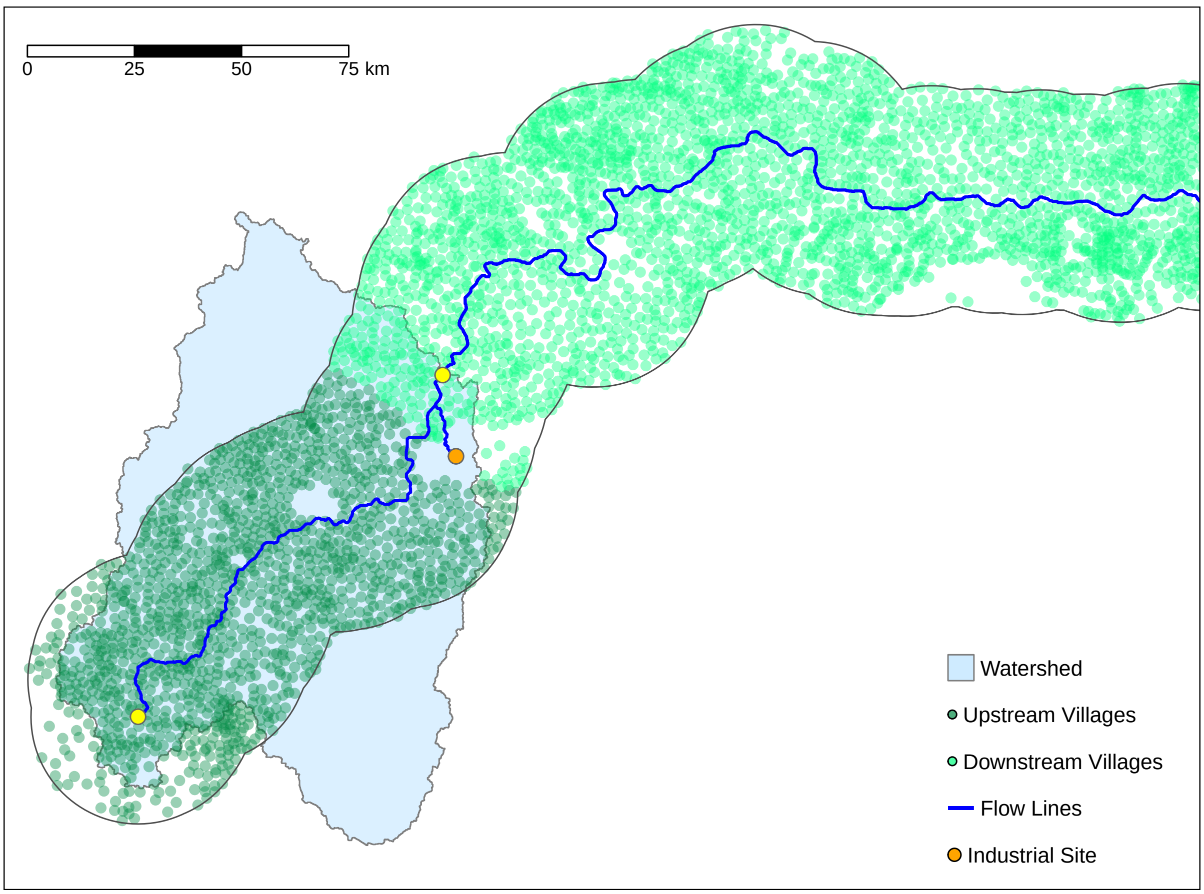


Figure 4: Continuity tests for a selection of covariates.

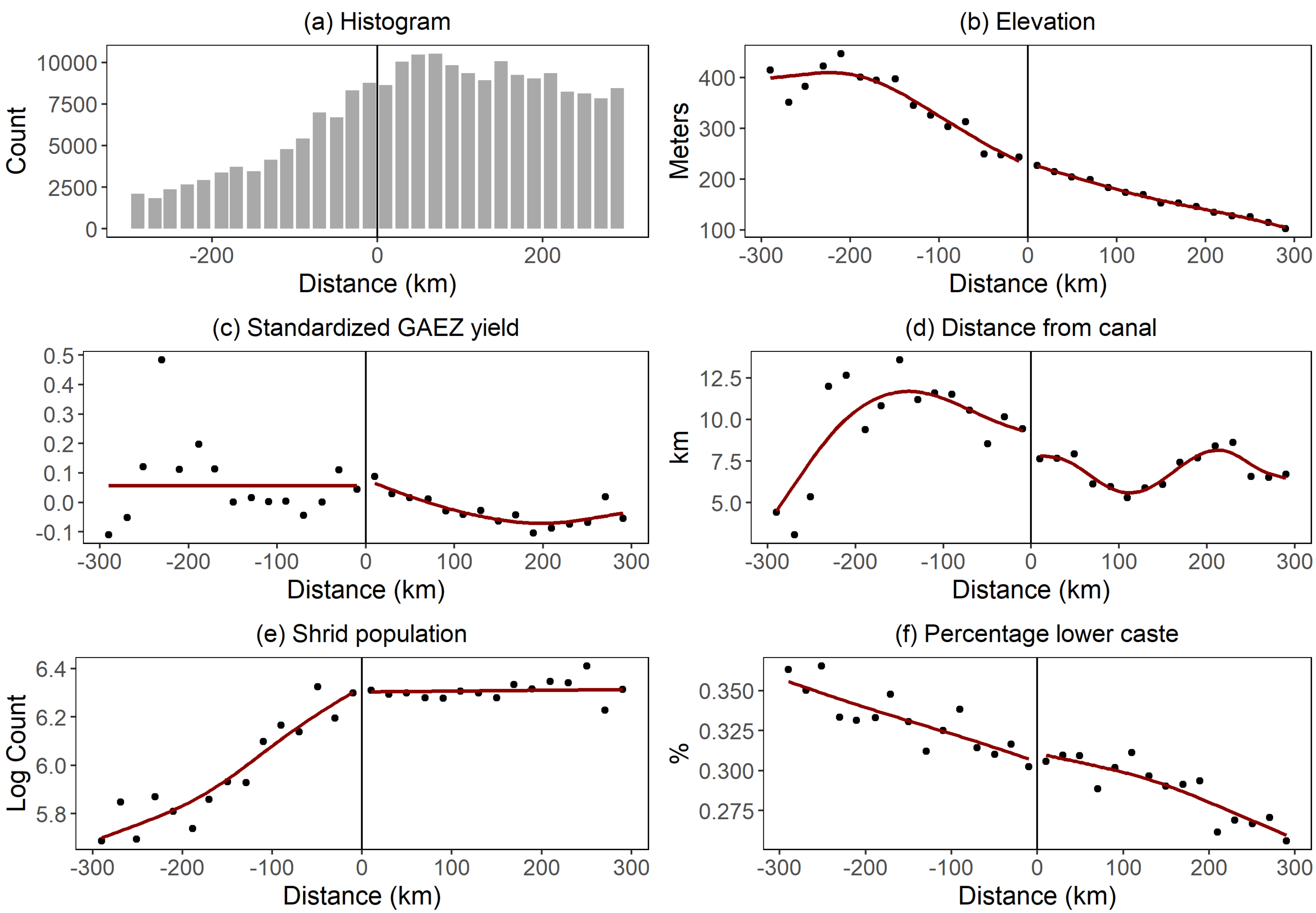


Figure 5: Regression discontinuity plots for pollution measurements. ↩

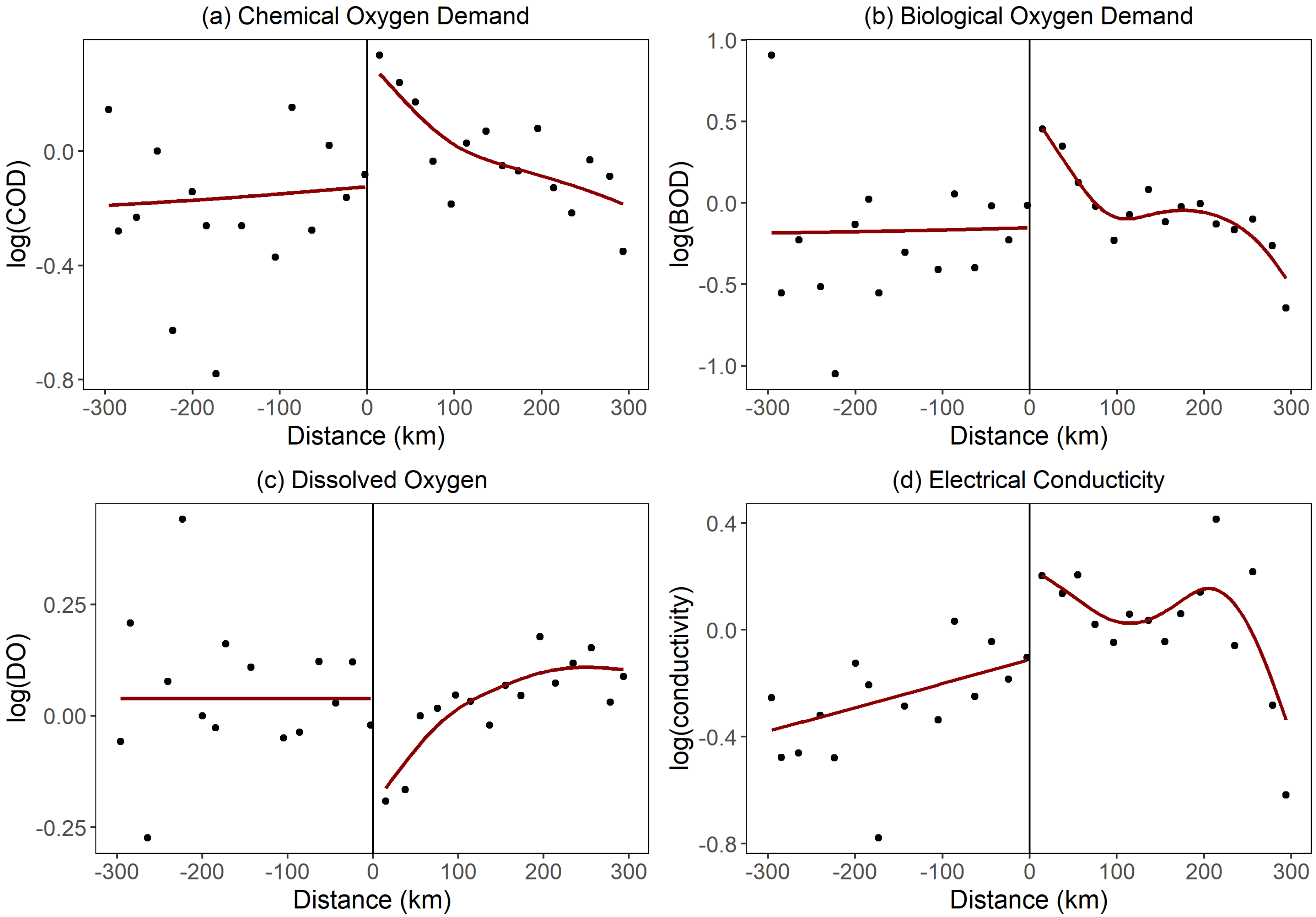


Figure 6: Regression discontinuity plots for measures of agricultural production. ↩

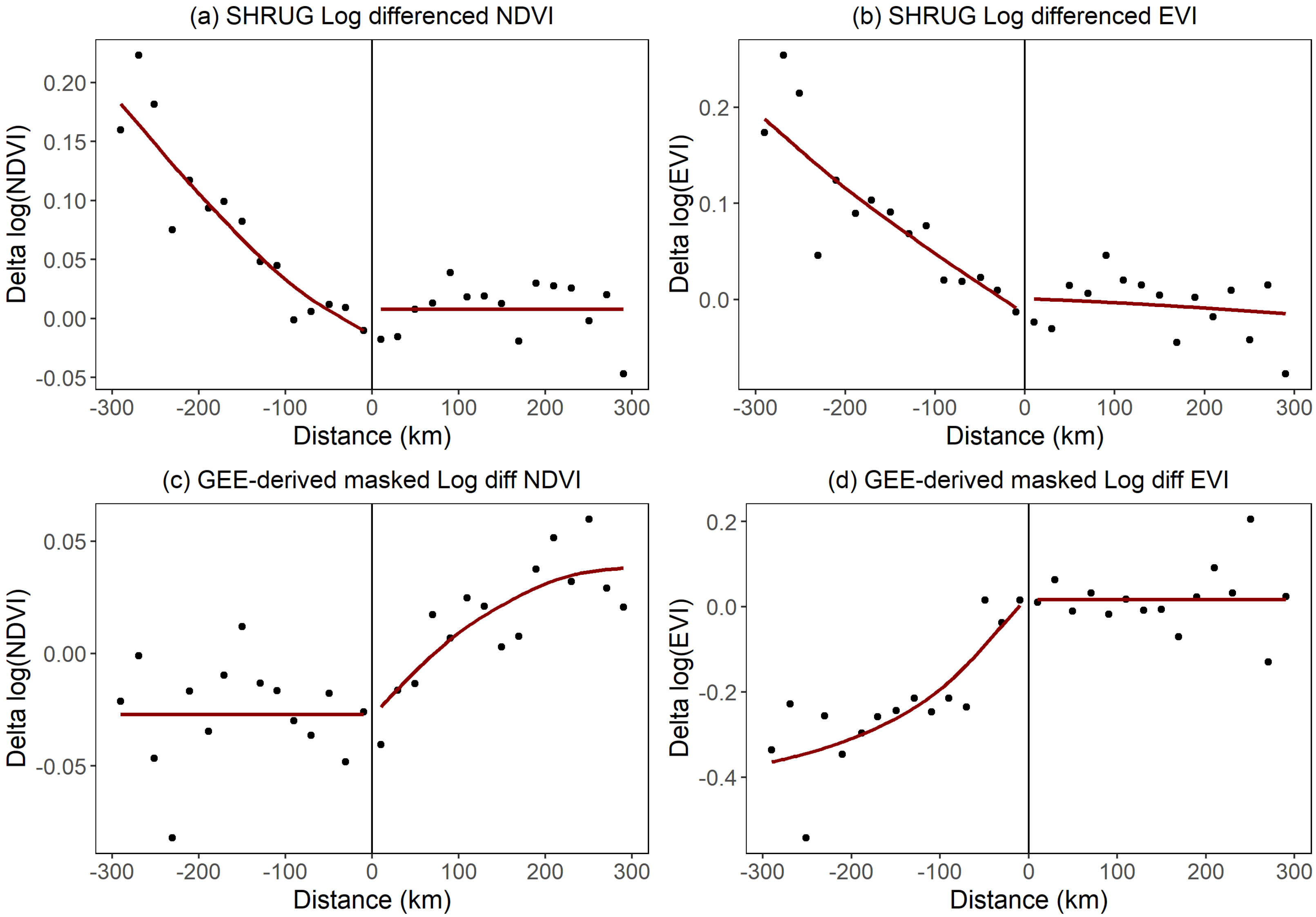


Table 1: Summary Statistics

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

continued

Table 1: Summary Statistics (Continued)

Notes: Summary statistics for the full sample of villages that are either upstream or downstream of severely-polluting industrial sites. Pollution data come from laboratory tests of samples taken at water quality monitoring stations maintained by the Central Pollution Control Board. NDVI and EVI variables from Google Earth Engine (GEE) are the mean of the log of each pixel's difference between maximum and minimum NDVI values within a year, for all village pixels marked as cropland in the cropland mask. NDVI and EVI variables from SHRUG are the log of the difference between the maximum and early-season village-mean NDVI values within a year, with no cropland mask applied. The net primary production (NPP) variable is the log of the mean of estimated NPP values across all cropland pixels within the village. Economic outcomes come from the Population Census of 2001. ↩

Table 2: RD Estimates for Pollution

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

continued

Table 2: RD Estimates for Pollution (Continued)

Notes: Estimated effects of severely-polluting industrial sites on water pollution concentrations in nearby rivers, immediately downstream of the sites. Dependent variables are listed in rows; each cell reports the estimated coefficient on the Downstream indicator variable, controlling linearly for distance on each side of the industrial site along with site-by-year fixed effects. Observations are limited to monitoring stations within the specified bandwidth of the industrial site and are weighted using a triangular kernel. Standard errors clustered by district. *p<0.1; **p<0.05; ***p<0.01.

Table 3: RD Estimates for Agricultural Outcomes

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

continued

Table 3: RD Estimates for Agricultural Outcomes (Continued)

Industry X Year FE	X	X	X

Notes: Estimated effects of severely-polluting industrial sites on remote sensing measures of agricultural production in villages immediately downstream of the sites. Dependent variables are listed in rows; each cell reports the estimated coefficient on the Downstream indicator variable, controlling linearly for distance on each side of the industrial site along with site-by-year fixed effects. Sample includes villages within 25 km of a flow path that passes near each industrial site, as defined in the text. Observations are limited to villages within the specified bandwidth of the industrial site and are weighted using a triangular kernel. Standard errors clustered by district. *p<0.1; **p<0.05; ***p<0.01. ↩

Table 4: RD Estimates for Agricultural Inputs and Economic Outcomes

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

continued

Table 4: RD Estimates for Agricultural Inputs and Economic Outcomes (Continued)

Industry FE	X	X	X

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

4.9 Appendix Tables

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

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

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

Table 6: RD Estimates for Continuity of Covariates

		RD Bandwidth		
Dependent Variable	Mean [SD]	[25 km]	[50 km]	[100 km]
Panel A: Infrastructure - Facility Available in Village?				
Banking	0.154	-0.033	-0.032	-0.021
	[0.361]	(0.021)	(0.020)	(0.016)
Communication	0.57	-0.004	0.008	0.014
	[0.495]	(0.024)	(0.020)	(0.018)
Medical	0.548	-0.008	-0.025	-0.023
	[0.498]	(0.033)	(0.036)	(0.033)
Postal	0.691	0.004	0.018	0.037**
	[0.462]	(0.021)	(0.015)	(0.015)
Paper and magazines	0.659	-0.083***	-0.015	0.015
	[0.474]	(0.032)	(0.022)	(0.022)
Educational	0.932	-0.003	-0.0004	0.005
	[0.252]	(0.007)	(0.006)	(0.008)
Drinking water	0.998	-0.0002	$1.64times10^{-5}$	0.0002
	[0.048]	(0.002)	(0.002)	(0.002)
Panel B: Physical Characteristics				
Distance from canal (km)	7.849	-0.426	-0.523	-1.24
	[11.585]	(0.833)	(0.761)	(0.794)
Distance from nearest town (km)	52.937	-2.22***	-1.65	-2.15*
	[549.917]	(0.847)	(1.53)	(1.09)
Elevation (m)	249.425	-3.54	-6.66**	-8.94**

continued

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

	[169.407]	(3.14)	(3.01)	(4.50)
Panel C: GAEZ potential yield - High Input Scenario				
Normalized All Crops	-0.224	-0.004	-0.008	-0.022
	[0.703]	(0.038)	(0.033)	(0.031)
Chickpea	0.585	-0.025	-0.025	-0.010
	[0.51]	(0.023)	(0.021)	(0.026)
Cotton	0.769	-0.002	-0.0002	-0.007
	[0.173]	(0.011)	(0.010)	(0.009)
Dryland rice	1.081	0.025	0.032	0.026
	[1.216]	(0.023)	(0.024)	(0.024)
Gram	1.476	-0.004	0.0009	-0.015
	[0.408]	(0.026)	(0.022)	(0.022)
Groundnut	1.404	-0.0003	-0.011	-0.024
	[0.497]	(0.026)	(0.022)	(0.023)
Maize	6.723	-0.036	-0.013	-0.058
	[1.975]	(0.115)	(0.099)	(0.102)
Pearl millet	1.263	0.008	0.024	0.027
	[1.321]	(0.028)	(0.026)	(0.029)
Pigeon pea	1.914	0.004	0.006	-0.015
	[0.632]	(0.033)	(0.029)	(0.028)
Rapeseed	0.854	0.013	0.005	0.009
	[0.655]	(0.019)	(0.016)	(0.017)
Sorghum	5.917	-0.052	-0.038	-0.045

continued

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

	[1.35]	(0.100)	(0.086)	(0.083)
Soybean	2.12	0.020	0.027	0.001
	[0.745]	(0.042)	(0.038)	(0.034)
Sugarcane	1.179	-0.005	0.023	0.031
	[1.866]	(0.027)	(0.043)	(0.056)
Sunflower	1.029	-0.004	-0.060*	-0.079*
	[0.724]	(0.023)	(0.035)	(0.044)
Wetland rice	1.727	-0.012	-0.019	-0.023
	[1.091]	(0.037)	(0.037)	(0.046)
Wheat	1.321	0.002	-0.016	-0.026
	[1.141]	(0.034)	(0.032)	(0.032)
Panel D: Social and Demographic Characteristics				
Household size	5.785	0.119***	0.076*	0.024
	[0.931]	(0.043)	(0.041)	(0.040)
Literacy Rate (percent)	0.503	-0.009	-0.003	0.003
	[0.133]	(0.009)	(0.007)	(0.007)
Log Village Area	6.281	-0.065	-0.051	-0.025
	[1.035]	(0.055)	(0.056)	(0.047)
Log Population	7.455	-0.062	-0.032	-0.040
	[1.056]	(0.050)	(0.044)	(0.038)
Share of Scheduled Caste/Tribe Population	0.283	-0.016	-0.005	0.0005
	[0.228]	(0.018)	(0.014)	(0.012)
Observations	85,745	22,364	44,982	85,745

continued

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

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

Notes: Tests of continuity in river space at severely-polluting industrial sites, for covariates that are either fixed in time or unlikely to be affected by the presence of industrial pollution. Each cell reports a regression discontinuity (RD) coefficient using regressions as described in Table 3. *p<0.1; **p<0.05; ***p<0.01. ↩

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

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

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

Bibliography

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

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. 2015. "The Economics of Density: Evidence From the Berlin Wall." *Econometrica : Journal of the Econometric Society* 83 (6): 2127–89.

Andini, Ade, Sébastien Bonnet, Patrick Rousset, and Udin Hasanudin. 2018. "Impact of Open Burning of Crop Residues on Air Pollution and Climate Change in Indonesia." *Current Science* 115 (12): 2259–66.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11 (1): 727–53.

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

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

Arrow, Kenneth. 1962. "Economic Welfare and the Allocation of Resources for Invention." In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, 609–26. Princeton University Press.

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

Assunção, Juliano, Clarissa Gandour, and Romero Rocha. 2022. "DETERring Deforestation in the Amazon: Environmental Monitoring and Law Enforcement."

American Economic Journal: Applied Economics.

Auffhammer, Maximilian, and Tamma A. Carleton. 2018. "Regional Crop Diversity and Weather Shocks in India." *Asian Development Review* 35 (2): 113–30.

Auffhammer, Maximilian, Solomon M. Hsiang, Wolfram Schlenker, and Adam Sobel. 2013. "Using Weather Data and Climate Model Output in Economic Analyses of Climate Change." *Review of Environmental Economics and Policy* 7 (2): 181–98.

Azam, Mehtabul. 2012. "The Impact of Indian Job Guarantee Scheme on Labor Market Outcomes: Evidence from a Natural Experiment." {IZA} working paper 6548.

Balboni, Clare. 2021. "In Harm's Way? Infrastructure Investments and the Persistence of Coastal Cities." Unpublished manuscript.

Banzhaf, H. Spencer, and B. Andrew Chupp. 2012. "Fiscal Federalism and Interjurisdictional Externalities: New Results and an Application to US Air Pollution." *Journal of Public Economics* 96 (5): 449–64.

Banzhaf, H. Spencer, and Randall P. Walsh. 2008. "Do People Vote with Their Feet? An Empirical Test of Tiebout." *American Economic Review* 98 (3): 843–63.

Barberis, Nicholas, and Ming Huang. 2008. "Stocks as Lotteries: The Implications of Probability Weighting for Security Prices." *American Economic Review* 98 (5): 2066–2100.

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

Behrer, A. Patrick. 2019. "The Impact of Anti-Poverty Efforts on Indian Agriculture and Air Pollution." Unpublished manuscript.

Bellemare, Marc F., and Casey J. Wichman. 2020. "Elasticities and the Inverse Hyperbolic Sine Transformation." *Oxford Bulletin of Economics and Statistics*

82 (1): 50–61.

Berg, Erlend, Sambit Bhattacharyya, D. Rajasekhar, and R. Manjula. 2018. "Can Public Works Increase Equilibrium Wages? Evidence from India's National Rural Employment Guarantee." *World Development* 103 (March): 239–54.

Berge, Laurent, Sebastian Krantz, and Grant McDermott. 2022. "Fixest: Fast Fixed-Effects Estimations."

Bertrand, Marianne, Bruno Crepon, Alicia Marguerie, and Patrick Premand. 2021. "Do Workfare Programs Live Up to Their Promises? Experimental Evidence from Côte d'Ivoire." Policy research working paper; no. 9611. Washington, DC: World Bank.

Besley, Timothy, and Stephen Coate. 1992. "Workfare Versus Welfare: Incentive Arguments for Work Requirements in Poverty-Alleviation Programs." *The American Economic Review* 82 (1): 249–61.

Besley, Timothy, and Torsten Persson. 2009. "The Origins of State Capacity: Property Rights, Taxation, and Politics." *American Economic Review* 99 (4): 1218–44.

Bhargava, Anil K. 2013. "The Impact of India's Rural Employment Guarantee on Demand for Agricultural Technology." 150163. Agricultural; Applied Economics Association.

Blackburn, Mckinley L. 2015. "The Relative Performance of Poisson and Negative Binomial Regression Estimators." *Oxford Bulletin of Economics and Statistics* 77 (4): 605–16.

Blakeslee, David, Ram Fishman, and Veena Srinivasan. 2020. "Way Down in the Hole: Adaptation to Long-Term Water Loss in Rural India." *American Economic Review* 110 (1): 200–224.

Bondy, Malvina, Sefi Roth, and Lutz Sager. 2020. "Crime Is in the Air: The Contemporaneous Relationship Between Air Pollution and Crime." *Journal of* *the Association of Environmental and Resource Economists* 7 (3): 555–85.

Borgschulte, Mark, David Molitor, and Eric Zou. 2022. "Air Pollution and the Labor Market: Evidence from Wildfire Smoke." {NBER} working paper 29952. Working Paper Series. National Bureau of Economic Research.

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.

Bryan, Gharad, and Melanie Morten. 2019. "The Aggregate Productivity Effects of Internal Migration: Evidence from Indonesia." *Journal of Political Economy* 127 (5): 2229–68.

Burgess, Robin, Francisco J M Costa, and Ben Olken. 2019. "The Brazilian Amazon's Double Reversal of Fortune." Preprint. SocArXiv.

Burgess, Robin, Olivier Deschenes, Dave Donaldson, and Michael Greenstone. 2017. "Weather , Climate Change and Death in India." Unpublished manuscript.

Burke, Marshall, Anne Driscoll, Sam Heft-Neal, Jiani Xue, Jennifer Burney, and Michael Wara. 2021. "The Changing Risk and Burden of Wildfire in the United States." *Proceedings of the National Academy of Sciences* 118 (2): e2011048118.

Burke, Marshall, and Kyle Emerick. 2016. "Adaptation to Climate Change: Evidence from US Agriculture." *American Economic Journal: Economic Policy* 8 (3): 106–40.

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.

Bustos, Paula, Bruno Caprettini, and Jacopo Ponticelli. 2016. "Agricultural Productivity and Structural Transformation: Evidence from Brazil." *American*

Economic Review 106 (6): 1320–65.

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–230.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2020. "Optimal Bandwidth Choice for Robust Bias Corrected Inference in Regression Discontinuity Designs." 1809.00236. arXiv.org.

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

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

Chang, Tom Y., Joshua Graff Zivin, Tal Gross, and Matthew Neidell. 2019. "The Effect of Pollution on Worker Productivity: Evidence from Call Center Workers in China." *American Economic Journal: Applied Economics* 11 (1): 151–72.

Chauvin, Juan Pablo, Edward Glaeser, Yueran Ma, and Kristina Tobio. 2016. "What Is Different About Urbanization in Rich and Poor Countries? Cities in Brazil, China, India and the United States." Working {Paper} 22002. National Bureau of Economic Research.

Chen, Shuai, Paulina Oliva, and Peng Zhang. 2022. "The Effect of Air Pollution on Migration: Evidence from China." *Journal of Development Economics* 156 (May): 102833.

Chen, Yuyu, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. 2013. "Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China's Huai River Policy." *Proceedings of the National Academy of Sciences of the United States of America* 110 (32): 12936–41.

Cole, Shawn A., and Wentao Xiong. 2017. "Agricultural Insurance and Economic Development." *Annual Review of Economics* 9 (1): 235–62.

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

Costinot, Arnaud, Dave Donaldson, and Cory Smith. 2016. "Evolving Comparative Advantage and the Impact of Climate Change in Agricultural Markets: Evidence from 1.7 Million Fields Around the World." *Journal of Political Economy* 124 (1): 205–48.

D'Agostino, Anthony Louis, and Wolfram Schlenker. 2016. "Recent Weather Fluctuations and Agricultural Yields: Implications for Climate Change." *Agricultural Economics* 47 (S1): 159–71.

Dechezleprêtre, Antoine, Nicholas Rivers, and Balazs Stadler. 2019. "The Economic Cost of Air Pollution: Evidence from Europe." {OECD} {Working} paper 1584. Paris.

Deininger, Klaus, Hari K. Nagarajan, and Sudhir K. Singh. 2016. "Short-Term Effects of India's Employment Guarantee Program on Labor Markets and Agricultural Productivity." Policy {Research} {Working} {Paper} {No}. 7665. World Bank, Washington DC.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. 2012. "Temperature Shocks and Economic Growth: Evidence from the Last Half Century." *American Economic Journal: Macroeconomics* 4 (3): 66–95.

Dercon, Stefan. 2002. "Income Risk, Coping Strategies, and Safety Nets." *World Bank Research Observer* 17 (2): 141–66.

Deryugina, Tatyana, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif. 2019. "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction." *American Economic Review* 109 (12): 4178–4219.

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

Down To Earth. 1999. "Parched Punjab." *Parched Punjab*. https://www.downtoearth.org.in/coverage/parched-punjab-20535.

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.

———. 2018. "The Value of Regulatory Discretion: Estimates From Environmental Inspections in India." *Econometrica* 86 (6): 2123–60.

Duflo, Esther, and Rohini Pande. 2007. "Dams*." *The Quarterly Journal of Economics* 122 (2): 601–46.

Duranton, Gilles, Ejaz Ghani, Arti Grover Goswami, and William Kerr. 2016. "A Detailed Anatomy of Factor Misallocation in India." Policy {Research} {Working} {Paper};{No}. 7547. Washington, DC: World Bank.

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

Dutta, Puja, Rinku Murgai, Martin Ravallion, and Dominique Van de Walle. 2012. *Does India's Employment Guarantee Scheme Guarantee Employment?* Policy Research Working Papers. The World Bank.

Fajgelbaum, Pablo D, Eduardo Morales, Juan Carlos Suárez Serrato, and Owen Zidar. 2019. "State Taxes and Spatial Misallocation." *The Review of Economic Studies* 86 (1): 333–76.

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

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

Fishman, Ram. 2018. "Groundwater Depletion Limits the Scope for Adaptation to Increased Rainfall Variability in India." *Climatic Change* 147 (1): 195–209.

Foster, Andrew D., and Mark R. Rosenzweig. 2017. "Are There Too Many Farms in the World? Labor-Market Transaction Costs, Machine Capacities and Optimal Farm Size." Working {Paper} 23909. National Bureau of Economic Research.

Freeman, Richard, Wenquan Liang, Ran Song, and Christopher Timmins. 2017. "Willingness to Pay for Clean Air in China." Working paper 24157. National Bureau of Economic Research.

Fu, Shihe, V Brian Viard, and Peng Zhang. 2021. "Air Pollution and Manufacturing Firm Productivity: Nationwide Estimates for China." *The Economic Journal* 131 (640): 3241–73.

Gehrke, Esther. 2019. "An Employment Guarantee as Risk Insurance? Assessing the Effects of the NREGS on Agricultural Production Decisions." *World Bank Economic Review* 33 (2): 413–35.

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.

Ghanem, Dalia, and Junjie Zhang. 2014. "'Effortless Perfection:' Do Chinese Cities Manipulate Air Pollution Data?" *Journal of Environmental Economics and Management* 68 (2): 203–25.

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

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

Graff Zivin, Joshua, and Matthew Neidell. 2012. "The Impact of Pollution on Worker Productivity." *American Economic Review* 102 (7): 3652–73.

Greenstone, Michael. 2021. "AQLI India Fact Sheet."

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

Greenstone, Michael, Guojun He, Ruixue Jia, and Tong Liu. 2022. "Can Technology Solve the Principal-Agent Problem? Evidence from China's War on Air Pollution." *American Economic Review: Insights*, April.

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

Grossman, Gene M., and Alan B. Krueger. 1995. "Economic Growth and the Environment*." *The Quarterly Journal of Economics* 110 (2): 353–77.

Hagerty, Nick. 2021. "Adaptation to Surface Water Scarcity in Irrigated Agriculture." Unpublished manuscript.

Hallegatte, Stephane, Mook Bangalore, Laura Bonzanigo, Marianne Fay, Tamaro Kane, Ulf Narloch, Julie Rozenberg, David Treguer, and Adrien Vogt-Schilb. 2016. *Shock Waves: Managing the Impacts of Climate Change on Poverty*. Washington, DC: World Bank.

Hammer, Melanie S., Aaron van Donkelaar, Chi Li, Alexei Lyapustin, Andrew M. Sayer, N. Christina Hsu, Robert C. Levy, et al. 2020. "Global Estimates and Long-Term Trends of Fine Particulate Matter Concentrations (1998–2018)." *Environmental Science & Technology* 54 (13): 7879–90.

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

Heblich, Stephan, Alex Trew, and Yanos Zylberberg. 2021. "East-Side Story: Historical Pollution and Persistent Neighborhood Sorting." *Journal of Political Economy* 129 (5): 1508–52.

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

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India*." *The Quarterly Journal of Economics* 124 (4): 1403–48.

Hsieh, Chang-Tai, and Enrico Moretti. 2019. "Housing Constraints and Spatial Misallocation." *American Economic Journal: Macroeconomics* 11 (2): 1–39.

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*.

Imbert, Clément, and John Papp. 2015. "Labor Market Effects of Social Programs: Evidence from India's Employment Guarantee." *American Economic Journal: Applied Economics* 7 (2): 233–63.

Jack, B Kelsey, Seema Jayachandran, Namrata Kala, and Rohini Pande. 2022. "Money (Not) to Burn: Payments for Ecosystem Services to Reduce Crop Residue Burning." Unpublished manuscript.

Jayachandran, Seema. 2006. "Selling Labor Low: Wage Responses to Productivity Shocks in Developing Countries." *Journal of Political Economy* 114 (3): 538–75.

———. 2009. “Air Quality and Early-Life Mortality Evidence from Indonesia’s Wildfires.” *Journal of Human Resources* 44 (4).

———. 2022. “How Economic Development Influences the Environment.” *Annuul Review of Economics* 14 (September): 1–30.

Kahn, Matthew E., Pei Li, and Daxuan Zhao. 2015. “Water Pollution Progress at Borders: The Role of Changes in China’s Political Promotion Incentives.” *American Economic Journal: Economic Policy* 7 (4): 223–42.

Kataki, P. K., P. Hobbs, and B. Adhikary. 2001. “The Rice-Wheat Cropping System of South Asia.” *Journal of Crop Production* 3 (2): 1–26.

Kaur, Supreet. 2019. “Nominal Wage Rigidity in Village Labor Markets.” *American Economic Review* 109 (10): 3585–3616.

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

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

Khanna, Gaurav, Wenquan Liang, A. Mushfiq Mobarak, and Ran Song. 2021. “The Productivity Consequences of Pollution-Induced Migration in China.” *Discussion Papers*, January.

Khera, Reetika, ed. 2011. *The Battle for Employment Guarantee*. Oxford, New York: Oxford University Press.

Kurukulasuriya, Pradeep, and Robert Mendelsohn. 2008. “Crop Switching as a Strategy for Adapting to Climate Change.” *African Journal of Agricultural and Resource Economics* 02 (1): 1–22.

Lin, Wei, and Jeffrey M. Wooldridge. 2019. “Chapter 2 - Testing and Correcting for Endogeneity in Nonlinear Unobserved Effects Models.” In *Panel Data Econometrics*, edited by Mike Tsionas, 21–43. Academic Press.

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

Lipscomb, Molly, and Ahmed Mushfiq Mobarak. 2017. "Decentralization and Pollution Spillovers: Evidence from the Re-Drawing of County Borders in Brazil*." *The Review of Economic Studies* 84 (1): 464–502.

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

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

Lohan, Shiv Kumar, H. S. Jat, Arvind Kumar Yadav, H. S. Sidhu, M. L. Jat, Madhu Choudhary, Jyotsna Kiran Peter, and P. C. Sharma. 2018. "Burning Issues of Paddy Residue Management in North-West States of India." *Renewable and Sustainable Energy Reviews* 81 (January): 693–706.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142 (2): 698–714.

McDonald, Andrew J., Balwinder-Singh, Alwin Keil, Amit Srivastava, Peter Craufurd, Avinash Kishore, Virender Kumar, et al. 2022. "Time Management Governs Climate Resilience and Productivity in the Coupled Rice–Wheat Cropping Systems of Eastern India." *Nature Food* 3 (7): 542–51.

Mitton, Todd, and Keith Vorkink. 2007. "Equilibrium Underdiversification and the Preference for Skewness." *The Review of Financial Studies* 20 (4): 1255–

88.

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

Morduch, Jonathan. 1995. "Income Smoothing and Consumption Smoothing." *Journal of Economic Perspectives* 9 (3): 103–14.

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

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2016. "Building State Capacity: Evidence from Biometric Smartcards in India." *American Economic Review* 106 (10): 2895–2929.

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

Olea, José Luis Montiel, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics* 31 (3): 358–69.

Parikh, Kirit S., A. Ganesh-Kumar, and Gangadhar Darbha. 2003. "Growth and Welfare Consequences of Rise in MSP." *Economic and Political Weekly* 38 (9): 891–95.

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

Ranson, Matthew. 2014. "Crime, Weather, and Climate Change." *Journal of Environmental Economics and Management* 67 (3): 274–302.

Ravallion, Martin. 1991. "REACHING THE RURAL POOR THROUGH PUBLIC EMPLOYMENT: Arguments, Evidence, and Lessons from South Asia." *The World Bank Research Observer* 6 (2): 153–75.

Redding, Stephen, and Esteban Rossi-Hansberg. 2017. "Quantitative Spatial Economics." *Annual Review of Economics* 9 (1): 21–58.

Redding, Stephen, and Daniel M. Sturm. 2008. “The Costs of Remoteness: Evidence from German Division and Reunification.” *American Economic Review* 98 (5): 1766–97.

Redding, Stephen, and Matthew Turner. 2015. “Transportation Costs and the Spatial Organization of Economic Activity.” In *Handbook of Regional and Urban Economics*, 1339–98. Elsevier.

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

Rigolini, Jamele. 2021. “Social Protection and Labor: A Key Enabler for Climate Change Adaptation and Mitigation.” Social {Protection} and {Jobs} {Discussion} {Paper} 2108. Washington, DC: World Bank.

Roback, Jennifer. 1982. “Wages, Rents, and the Quality of Life.” *Journal of Political Economy* 90 (6): 1257–78.

Rodrik, Dani. 2008. “Second-Best Institutions.” *American Economic Review* 98 (2): 100–104.

Rogers, Haley M., Jenna C. Ditto, and Drew R. Gentner. 2020. “Evidence for Impacts on Surface-Level Air Quality in the Northeastern US from Long-Distance Transport of Smoke from North American Fires During the Long Island Sound Tropospheric Ozone Study (LISTOS) 2018.” *Atmospheric Chemistry and Physics* 20 (2): 671–82.

Rosen, Sherwin. 1974. “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.” *Journal of Political Economy* 82 (1): 34–55.

Rosenzweig, Mark R., and Christopher Udry. 2014. “Rainfall Forecasts, Weather, and Wages over the Agricultural Production Cycle.” *American Economic Review* 104 (5): 278–83.

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.

Ryan, Nick, and Anant Sudarshan. 2020. "Do India's Farmers Use Too Much Water?" {VoxDev}.

Santangelo, G. 2019. "Firms and Farms: The Local Effects of Farm Income on Firms' Demand." Working paper. University of Cambridge Faculty of Economics.

Schlenker, Wolfram, and W. Reed Walker. 2016. "Airports, Air Pollution, and Contemporaneous Health." *The Review of Economic Studies* 83 (2): 768–809.

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

Seo, S. Niggol, and Robert Mendelsohn. 2008. "An Analysis of Crop Choice: Adapting to Climate Change in South American Farms." *Ecological Economics* 67 (1): 109–16.

Sheahan, Megan, Yanyan Liu, Sudha Narayan, and Christopher B. Barrett. 2020. "Disaggregated Labor Supply Implications of Guaranteed Employment in India." Unpublished manuscript.

Shyamsundar, P., N. P. Springer, H. Tallis, S. Polasky, M. L. Jat, H. S. Sidhu, P. P. Krishnapriya, et al. 2019. "Fields on Fire: Alternatives to Crop Residue Burning in India." *Science (New York, N.Y.)*, August.

Singh, Karam. 2009. "Act to Save Groundwater in Punjab: Its Impact on Water Table, Electricity Subsidy and Environment." *Agricultural Economics Research Review* 22 (January).

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.

Stern, David. 2017. "The Environmental Kuznets Curve After 25 Years." *Journal of Bioeconomics* 19 (1): 7–28.

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by Donald W. K. Andrews and James H. Stock, 80–108. Cambridge: Cambridge University Press.

Subramanian, Kapil. 2017. "How Delhi's Air Pollution Crisis May Have Origins in the Green Revolution." *The Hindu*, November.

Sukhtankar, Sandip. 2017. "India's National Rural Employment Guarantee Scheme: What Do We Really Know about the World's Largest Workfare Program?" In *India Policy Forum*, 13:231–85.

Suri, Tavneet. 2011. "Selection and Comparative Advantage in Technology Adoption." *Econometrica* 79 (1): 159–209.

Taraz, Vis. 2017. "Adaptation to Climate Change: Historical Evidence from the Indian Monsoon." *Environment and Development Economics* 22 (5): 517–45.

———. 2018. "Can Farmers Adapt to Higher Temperatures? Evidence from India." *World Development* 112 (December): 205–19.

———. 2021. "Public Works Programs and Agricultural Risk: Evidence from India." Unpublished manuscript.

Tombe, Trevor, and Xiaodong Zhu. 2019. "Trade, Migration, and Productivity: A Quantitative Analysis of China." *American Economic Review* 109 (5): 1843–72.

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

Vallero, Daniel. 2014. *Fundamentals of Air Pollution - 5th Edition*. 5th ed. Elsevier.

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

Wang, Jinxia, Robert Mendelsohn, Ariel Dinar, and Jikun Huang. 2010. "How Chinese Farmers Change Crop Choice to Adapt to Climate Change." *Climate Change Economics* 01 (03): 167–85.

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

Wooldridge, Jeffrey M. 1999a. "Quasi-Likelihood Methods for Count Data." In *Handbook of Applied Econometrics Volume 2: Microeconomics*, 321–68. John Wiley & Sons, Ltd.

———. 1999b. "Distribution-Free Estimation of Some Nonlinear Panel Data Models." *Journal of Econometrics* 90 (1): 77–97.

———. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA, USA: MIT Press.

World Bank. 2018. "Power Sector Distortions Cost India Billions." *World Bank*.

———. 2021. "Addressing Groundwater Depletion: Lessons from India, the World's Largest User of Groundwater."

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

Young, Nicholas E., Ryan S. Anderson, Stephen M. Chignell, Anthony G. Vorster, Rick Lawrence, and Paul H. Evangelista. 2017. "A Survival Guide to Landsat Preprocessing." *Ecology* 98 (4): 920–32.

Zhang, Yapei. 2018. "Seeking Skewness." Unpublished manuscript.