Environmental Shocks in Ghana: An Improved Detection of their Impact on Child Health

Vidisha Vachharajani*

University of Illinois, Urbana-Champaign

Job Market Paper

Draft: Do not cite without permission.

Abstract

This paper uses Ghanaian household survey data to examine the link between environmental factors and the long-term health status of children in Western Ghana by exploiting the variation induced by a mining-based cyanide spill into a major water body. Information in the survey about the region and cohort of birth gives the primary source of identification. However, to get an improved spatial identification of exposed children, I link GPS data to the survey data. A rich set of controls are included to allow for a better resolution of omitted variable bias, by accounting for variables that can potentially confound the impact estimate if left unidentified. I also examine quantile treatment effects using a flexible specification, which allows for uniquely incorporating the GPS information. Findings reveal that after controlling for birth region and cohort, household, maternal and environmental factors, children born during the shock in the WR are negatively affected with reduced height, and that this negative effect persists through all the above (baseline and alternative) specifications. This effect initially reduces in magnitude, as we begin adding controls, but stabilizes across specifications once we include a larger set of control variables. A discussion motivated by the appropriate clustering of standard errors of the shock impact coefficient is presented.

Keywords: Child health, environment, chemical spills, Africa, additive models, quantile regression.
JEL Code: C14, C21, C52, I10, J13, O12, O15.

*Address: 214 David Kinley Hall, 1407 W. Gregory Dr., Urbana IL 61801, email: *vvachha2@illinois.edu*.
I am grateful to Anil Bera, Roger Koenker, Richard Akresh, Darren Lubotsky, Elizabeth Powers, and Dan Bernhardt for very helpful comments. All errors are mine.

1 Introduction

"*We are drowning in information and starving for knowledge.*"

Rutherford D. Roger

What are the health consequences of an interaction of factors from the natural and the socioeconomic environment? In this paper, I examine the impact of environmental and mining shocks on child health in Ghana. Since mining forms the basis of many livelihoods in the Western Region (WR), it is crucial to link these shocks to health outcomes. Early childhood malnutrition has far-extending consequences, including outcomes like educational attainment and earnings. There is ample evidence linking early childhood and health outcomes later in life (see reviews like Godfrey and J.P. (2000) and Alderman and Behrman (2006)). Using Netherlands' birth registry data, Van den Berg et al. (2006) determine that poor macroeconomic conditions in infancy lead to higher adult mortality. Findings in Almond et al. (2007) suggest that fetal exposure to acute maternal malnutrition compromises literacy, labor market, wealth, and marriage outcomes. Akresh et al. (2007) examine the effects of exogenous shocks like civil war and crop failure on childhood stunting.

Existing literature examining shock impact on health usually considers large scale events. For example, in Bundervoet et al. (2009), several provinces in Burundi were affected by civil war, making it a wide-spread national event. In this paper, I focus on chemical (cyanide) spill on account of mining activity that affects only a section of the WR in 2001. I use data from the 2003 Demographic and Health Survey (DHS) for children under age 5. However, the 'treated' population is concentrated in one area within the affected state, and our survey data is across states/regions. Although it can be argued that the impact is not strictly localized to the spill site, we need to exclude sections of the region that are obviously unaffected. This paper contributes to existing child health literature by exploiting information in the Global Positioning System (GPS) data associated with the 2003 survey. This contains information for children at their enumeration area (EA) or 'village' level, which enables greater spatial variation in the estimation by controlling for the child's birth (sampling) cluster, which

consists of a group of households, rather than birth region. It also critically examines the nature of alternative specifications across which we expect a changing magnitude of the shock impact, to identify the one that can best capture this effect.

For specifications that identify the whole of WR as shock-affected, I augment the models to better specify shock exposure. I interact relevant factors like gender, and water source with the shock exposure variable, to construct an intermediate link between the shock at birth and health outcome. This helps to measure the exposure intensity of the shock. Effects of these shocks on height-for-age are estimated after controlling for month and year of birth and additional controls like household, maternal and child characteristics.

In the process of exploring alternative and better specifications, another contribution of this paper is to extend the analysis of the shock impact ('treatment' effect) in 2 ways. First, I incorporate GPS information uniquely by expressing child height as a smooth bivariate function of latitude and longitude. This demonstrates benefits of smoothing on geographical location, a strategy being increasingly implemented within the generalized additive model (GAM) framework for improved *spatial* analyses in environmental health models (Vieira et al. (2010)). Second, I examine shock impact across quantiles of child height. I implement this by including exposure as a covariate in a nonparametric quantile regression model, where the flexible additive specification follows Koenker (2010). The three main components of Koenker's specification are: an additive nonparametric objective function, a (total variation) penalty for continuous (nonparametric) risk factors, and a LASSO penalty for categorical (linear) covariates, in the objective function. Control variables with a potentially nonlinear effect are included as smooth components, subject to penalty. The LASSO selection embedded in the dual penalty, allows for a second, more informative level of model selection, because it derives the impact of the shock at different points in the height distribution. I contrast the results across models that exclude and then include smoothed latitude-longitude components to allow for better control for the additional geographical information.

A final contribution of the paper is in exploring the robustness of the results by examining them along one other dimension: accuracy of the standard error of the shock impact

coefficient. In microeconometric applications, in addition to ensuring an unbiased estimate, the standard error should be estimated accurately to avoid an overly high rate of rejection of the null hypothesis of no effect. After a complete spatial identification, the variation in the 'treatment' variable is along the 'village' unit. Since we believe there are 'village' effects that cannot be observed, OLS-type procedures will produce inefficient standard errors. After including these village effects as dummies, we can then cluster standard errors along the region level. Finally, the disaggregation of the clustering unit is examined with respect to a caveat of re-estimating standard errors after accounting for the fact that the 'treatment' is across fewer units relative to the control units (Conley and Taber (2009)).

The findings are summarized below:

1. Primary results from the fixed effects models suggest that children exposed to the shock have a height-for-age reduced by 0.15 standard deviations. Although not very large, this estimate is larger than the one from a less accurate spatial identification of exposed children, where all children in the WR are considered as exposed. This suggests resolution of measurement error on account of noise in the exposure variable. Exposure intensity measures suggest that children in rural areas, and those exposed to natural/ground water sources are worse off. In contrast to typical findings in this literature, boys are actually worse off than girls.

2. Further, I examine the shock impact magnitude (size of the shock exposure coefficient) across specifications accounting for an increased level of unobserved spatial heterogeneity: from *only* within region variation to within EA-cluster variation. For models estimating the shock impact, with few additional control factors, the above specifications differ in the size of their shock exposure coefficient. This difference is caused by inadequate control for unobserved variation. However, when a systematically chosen, larger set of additional controls is used, this difference disappears, and the shock impact coefficient 'stabilizes' across these models. This alternative way of thinking about the 'omitted' effects in the shock impact estimation also removes any residual unobservable heterogeneity. This suggests an improved resolution of an omitted variable bias.

3. Using the quantile specification, I find a largely homogenous negative effect of the shock across all height quantiles. The impact is *more negative* in the lower tail than in the higher tail. More importantly, accounting for GPS coordinates in the specification, via smoothed effects of latitude and longitude, exhibits an increased negative impact than if we were to leave out this information. This strategy also uniquely resolves measurement error in the exposure variable.

These findings are relevant with respect to building an awareness about the consequences of an interaction of the natural environment with the livelihood source of a region. The paper is organized as follows. Section 2 gives a brief description of the Western Region in Ghana, especially pertaining to mining being its primary livelihood source. Section 3 describes the data and defines the health and shock variables used in the fixed effects models. Section 4 demonstrates the empirical identification strategy and summarizes the results. Section 5 examines the additive quantile regression model in context of the environmental shock impact. Section 6 concludes.

2 Western Region

There is little doubt about the strength of the mining industry in taking Ghana forward, especially gold mining. Ghana is the second-largest gold producer in West Africa. Its gold production was up 12% in 2009, with rising output from its 2 largest mines - Tarkwa and Ahafo. At the same time, there are large-scale regional and intra-region inequalities in economic and health outcomes across Ghana. The 3 regions in the north are much poorer than the rest of Ghana. As a result, these regions, especially the Northern Region (NR), became the focus point of the Ghana Poverty Reduction Scheme (GPRS) intervention. Regions consisting mostly of mining towns are conventionally perceived to be well-off, but tend to suffer ecological damage. Further, within mining industry-rich regions, there is a second level of inequality, stemming from the difference between the mainstream mining, and small-scale mining, the latter also known as the artisanal and small-scale mining (ASM)

sector, or more popularly, the *galamsey*.

The Western Region (WR) is located in the south-west of Ghana, bordered by Cote D'Ivoire on the west. Along with Brong Ahafo, it is a major part of Ghana's gold belt, and has attracted the highest number of explorations and mining companies. A typical example this is the Wassa West district, which contains the shock sites. It produces a majority of the gold exported by Ghana. Conventional criteria indicate that Wassa West should have a lower poverty level than rest of Ghana. It is said to have the highest agglomeration of mines and mining companies in Ghana, both large scale as well as ASM. It is also one of the most urbanized districts in the Western Region. Region hierarchy in Ghana comprises of: small towns, medium-sized/intermediate towns, large cities, and rural areas. The first 3 compose the urban centers, with large cities like Sekondi-Takoradi in the WR, intermediate towns being regional capitals, and finally the small towns being district capitals and administrative centers. Consider the map in Figure 2. The WR is composed of mining towns, the significant ones being Tarkwa and Prestea, all in the Wassa West district.

In October 2001, villages located within this triangle of mining towns in Wassa West were affected by a cyanide spill. The primary consequence was the loss of reliable water, which can likely be linked with malnutrition, especially when the affected water is a major source. Further, the spills are on a much lower scale as compared to the major ones like in Baia Mare, Romania in January 2000. There was an immediate clean-up by the concerning company and concentrations of the chemical in the water afterward were very low, although the water pollution had a likelihood of harming the immediate environment, aquatic life and potentially, human health. Nevertheless, the most probable channel of negative impact is via a shock to the closest water source, through a sudden change in its reliable and continuing availability. Table 1 summarizes some key health and development criteria across Ghanaian states. The WR has a high rural stunting rate in 2003, second only to the poorest region in Ghana, the Northern Region. It is one of the lowest in rural water supply availability in Ghana, which further decreases in 2008.

3 Data

3.1 Household Surveys

The Demographic and Health Survey (DHS) data consists of nationally-representative household surveys containing information on fertility, family planning, maternal and child health, child survival, HIV/AIDS, malaria, nutrition and anthropometric measures for children, like height, weight, and associated standardized Z-scores. In addition, there are demographic features for the child, parents and the household. The data I use in this paper has a sample of around 2700 children between age 0 and 5, surveyed in 2003. Detailed birth information for the child is also available, which is used for the primary identification strategy. The DHS also provides Global Positioning System (GPS) data, which specifies geographical latitude-longitude information for villages or enumeration areas (EAs) within each region, enabling a clearer spatial specification of exposed children. There are around 412 clusters across Ghana. I link this spatial data back to the survey data using these EA clusters. The GPS latitude-longitude coordinates are used for specifying the binary exposure variable in the fixed effects models, and then as smoothed components in the quantile specification for deriving the shock impact. They allow us to compute the distance between the shock site the villages, so that those closer to the site can be identified as being exposed.

3.2 Health and Environmental Shock Variables

Literature in child health suggests that an objective health measure is obtained by taking child height conditional on age and gender, and becomes a good long-run nutritional status measure, manifesting past deficits. I consider the height-for-age z-score (HAZ score) as the main health outcome. This is computed as the difference between the child's height and the median height of the same-aged international reference population, divided by the standard deviation of the reference population. Lower height-for-age indicates chronic malnutrition, which in developing countries is largely attributable to environmental factors. Table 2 summarizes this measure across Ghanaian states. On average, a Ghanaian child is 1.3 standard

deviations lower than the reference child. For instance, this translates to roughly 4 cm lower height for 2 year old girls. The poorest state in Ghana, Northern Region, has the lowest average height-for-age, and the capital city of Accra has the highest average height-for-age. We also see that the mean (or the median) child across all states is not stunted on average. Only children in the lower tails (below median) are stunted (moderately to severely).

Birth-related spatial (S) and temporal (T) variation is exploited to create the primary measure for shock exposure, which is used as a covariate to compute the estimated exposure impact. The structure of this variable is in the form of $S \times T$. The spatial specification considers either children born in the entire WR, or, exploiting GPS information, only those born in villages closely surrounding the spill site. The latter accounts for variation within the WR across children born *close to* and those born farther away from the shock site. The temporal specification is limited to children in the survey, born before and alive through the October 2001 shock, and those born right after, experiencing the shock *in utero*.

'Exposed' is defined as children born before and "during" the spill in WR, and non-exposed are those born after the spill in regions other than WR. Since the spill does not span a period of a few months, weeks or even days, "born during the spill" includes children born in the month of the spill and a few months afterward. Thus, I define exposure as a binary variable: *WR* $\times$ *Born During Spill*, allowing me to specify variation across regions, in terms of exposure to spill, and within region, in terms of timing of birth. I also define *WR* $\times$ *Alive through Spill*, which includes all children in the WR that are born before and are alive up until the spill occurs in 2001.

The GPS information allows a similar comparison, but allows for more within-WR variation by comparing children born closer to the site and those born away from it. This allows me to define: *Closer to Spill* $\times$ *Alive through Spill*. Using each village's latitude-longitude coordinates, I can identify villages closer to the spill to create the above variable. Since the spill occurs in the specified triangle in section 2, within the WR, this strategy can resolve measurement error on account of 'spatial noise' in the definition of exposure.

4 Identification Strategy

4.1 Nonparametric Evidence

With little information on the functional form of the child health production function, I gather preliminary evidence for the shocks from plots of nonparametric locally weighted regressions of the height-for-age on child age. This evidence illustrates the primary identification strategy. Figure 3 suggests declining heights-for-age for children born in 2001 in WR, as compared to those born elsewhere at the same time. The second vertical line indicates the end of the shock year. After specifying villages closer to the shock site using GPS information, Figure 4 shows how these villages, considered to be exposed, are worse off than those farther away from the shock. To expand on this evidence, I present box-plots of the Z-scores for children in 5, again for villages in the WR, either close or far from the spill, and the rest of Ghana. We see clearly how the whole distribution of height-for-age for villages closer to the spill site worsens closer to, and during 2001.

What is the impact of the shock across the height distribution? Figure 6 estimates height as a function of age for the extreme tails of the height distribution: $\tau = 0.2$ and $\tau = 0.8$, where τ is the height quantile. The 2 vertical lines mark the beginning and end of 2001. This figure suggests that the impact of the shock across all height quantiles is more or less homogenous across the distribution, so that children in both the upper and lower tails, born in 2001 in the WR, are worse off than children in the rest of Ghana.

Next, I uniquely use GPS information by expressing height-for-age as a smooth bivariate function of latitude and longitude. Under the generalized additive models (GAMs), the response and input variable can be linked nonlinearly using smooth functions, which can in turn be additively incorporated into the model. For instance, $y = s(x_2, x_1)$, where x_1 and x_2 are latitude and longitude respectively, and s is a smooth function. Shown in Figure 7, ech sub-figure can be seen as a geographical contour map, with the Latitude on the vertical axis, and the Longitude on the horizontal axis. Thus, it can be read as a map for the WR, with the black dots being the actual clusters or village. Further, temporal information is

incorporated, where the 4 figures correspond to the years, 1999, 2000, 2001 and 2002. Thus, each sub-figure has different number of clusters, depending on the number of children born in each year, and their location. In each one, the clusters are 'affected' based on the fact that they surround the river experiencing the spill, which runs from the north-east, from -2.1667 longitude, up till 5.3833 latitude, flows south to merge into the larger south-flowing river Ankobra. The spill occurred slightly upstream from the villages with coordinates (5.374, -2.016) and (5.382, -2.063), where the former, being more north, is closer to the site. I call these site 1 and site 2. Finally, the vertical bar on the right hand side of each plot is a black-and-white legend corresponding to height-for-ages across the cluster regions: darker is better off.

Figure 7(a) shows children born in 1999 by location. We see children who are 2-3 years old when the shock occurred are in the lighter gray region, with one cluster being in a lighter region, being further closer to the river. The clusters in the darker region are those in or surrounding Tarkwa, a major mining city, all of whom have a higher wealth index than those farther away from the city. In 2000, the lighter region covers a higher area, and the yellow region diminishes, indicating how younger children are more affected by the shock. The cluster closest to the first affected region is in the lighter region. This region covers a larger area for children born in 2001, with the clusters upstream, and close enough to where the spill occurs, also being in the blue region, and the one cluster being farther east and not close to the river, being in the green region. In 2002, all the affected regions are much better off. This strategy demonstrates the clarity of using GPS information via contour maps, and can be extremely beneficial for improved spatial analysis in a quasi-experimental setting, such as with an exogenous environmental shock.

4.2 DID and Fixed Effects

As an illustration of the identification strategy, I present a two-by-two difference-in-differences (DID) estimation summarized in Table 3. This uses the variable *WR* × *Born During Spill*, and considers the difference in height-for-age for children *born in* 2001 in the WR, relative to

those born in the rest of Ghana in 2001. Since it is a DID, the main coefficient should capture the difference in children born during the shock (in 2001) in the WR versus those in the rest of Ghana, over and above the difference between children *not* born in 2001 across these two. However, this excludes children born before 2001, and exposed to the shock, it not only demonstrates an inflated coefficient on the shock exposure variable, but also misspecification on account of not clearly identifying *non-exposed* children.

Panel A shows this coefficient to be significant, suggesting a decrease in height-for-age of these exposed children by 0.626 standard deviations. Boys are worse off than girls, with a lowered height-for-age by 0.994 standard deviations. This suggests there is some negative impact for children specified as exposed in this restricted manner. But the magnitude of this coefficient cannot be taken into account, since the non-exposed children have not been carefully identified, either spatially or temporally.

Thus, to augment these results, and to allow for appropriate variation, I estimate the following fixed-effects (FE) models:

$$\text{HAZ}_{ijt} = \mu_{1j} + \delta_t + \gamma_1 \times S_{1j} \times T_{1t} + X'_{ijt}\beta + \epsilon_{ijt} \tag{1a}$$

$$\text{HAZ}_{ijt} = \mu_{1j} + \delta_t + \gamma_2 \times S_{1j} \times T_{2t} + X'_{ijt}\beta + \epsilon_{ijt} \tag{1b}$$

$$\text{HAZ}_{ijt} = \mu_{1j} + \delta_t + \gamma_3 \times S_{2k} \times T_{2t} + X'_{ijt}\beta + \upsilon_{ijt} \tag{1c}$$

$$\text{HAZ}_{ijt} = \mu_{2k} + \delta_t + \gamma_4 \times S_{2k} \times T_{2t} + X'_{ijt}\beta + \vartheta_{ijt} \tag{1d}$$

where μ_{1j} is region fixed effects for regions $j = 1, \ldots, 10$, μ_{2k} is village fixed effects for villages $k = 1, \ldots, 412$, and δ_t is cohort fixed effects. Thus, equation (1a) estimates:

$$\text{HAZ}_{ijt} = \mu_1 + \delta_t + \gamma_1 \times \text{WR} \times \text{Born During Spill} + X'_{ijt}\beta + \epsilon_{ijt}, \tag{2}$$

the most general specification outlined in sub-section 3.2. Equation (1b) estimates:

$$\text{HAZ}_{ijt} = \mu_1 + \delta_t + \gamma_2 \times \text{WR} \times \text{Alive through Spill} + X'_{ijt}\beta + \epsilon_{ijt}. \tag{3}$$

Equations (1c) and (1d) use the variable *Closer to Spill* × *Alive through Spill* as the exposure variable, incorporating GPS information, with region and village FE respectively.

Note that this is not an actual panel. For each equation, the unit of measurement is the child i born in region j (or village k), in cohort t. The coefficient γ_i, $i = 1, 2, 3, 4$ measures the impact of the shock for different spatial and temporal indicator variables. We want to estimate γ effectively, by reducing misspecification stemming from measurement error and omitted variable bias. To equations (1a) and (1b), I add interactions like being a boy child, exposure to ground/surface water and residence type, to better specify exposure intensity in the absence of GPS information. Finally, I examine the change in magnitude of γ in the absence and presence of a rich set of controls, like child, maternal and socioeconomic characteristics.

4.3 Empirical Results

Tables 4 and 5 summarize results from equations (1). Table 4 focuses on the temporal definition *Born During Spill*, and 5 on *Alive through Spill*. Consider the baseline regressions in Table 4, estimating equation 2. Column (1) suggests that children born in the WR *during* the spill have a height-for-age lowered by 0.51 standard deviations. Columns (2) and (3) estimate the same, separately for rich and poor households. The definition of wealth is derived from the wealth quintiles specified in the DHS. We see how the poorer households are generally worse off, so that having bad cooking fuel affects them more. These households are also worse affected by the shock, with a higher coefficient magnitude. This supports the classic findings in the literature that socioeconomic background affects the response of households to exogenous shocks to income and health. Finally, columns (4) through (6) estimate exposure intensity of the shock using gender, water source and residence type as interacting factors. Boys are generally worse off relative to girls. Children born in rural areas, exposed to natural/ground water sources are worse off. All models include a rich set of additional control variables.

Although the above evidence generally supports the idea of the shock having a negative

impact on exposed children, the models are misspecified on account of including children from other parts of the WR. This could potentially inflate the coefficient, since there are pockets in the northern part of WR with very poor households that could be contributing to the coefficient magnitude.

Columns (1) through (3) in Table 5 estimates equation (3), which includes a baseline model, and 2 models with exposure intensity measures. In column (1), we see that on including all children born before and alive through the shock, the exposure variable specified by the WR is unable to capture any impact on height. Exposure intensity measures do demonstrate some negative effect of the shock. However, there is continued misspecification caused by measurement error, yielding an inaccurate exposure variable.

Finally, we incorporate GPS information to estimate the model which uses *Closer to Spill* $\times$ *Alive through Spill* as the exposure variable. Results are summarized in columns (4) through (7). Further, Columns (4) and (5) include "insufficient" number of additional controls in the specifying equations, so that we exclude important maternal and socioeconomic variables like mother's age, education, employment and health status, asset ownership and wealth variables, and others like child age. Columns (6) and (7) include this larger control set. Finally, columns (4) and (6) include region fixed effects, whereas (5) and (7) include village fixed effects, which controls for a higher degree of unobserved heterogeneity. The village FE models are appropriate on account of endogenous cluster effects of these villages, which could otherwise underestimate the standard errors if left unaccounted for, causing a false rejection of the null.

The following is observed: excluding relevant controls inflates the exposure coefficient on account of some confounding variable bias, and causes the magnitude to be different across models (4) and (5). Inclusion of all these controls 'stabilizes' the coefficient across (6) and (7). Based on more accurate temporal and spatial definition, specified by *Closer to Spill* $\times$ *Alive through Spill*, embedding GPS information in the exposure variable allows for capturing the impact of the shock. We see that children exposed to the spill have height-for-age reduced by 0.15 standard deviations. Comparing this to column (1)-(3) in the same table, we see how

the baseline exposure variable easily captures the negative (and arguably more accurate) impact of the spill, having included GPS coordinates in the exposure definition.

Invoking the motivation in Bertrand et al. (2004), I briefly consider the argument for a fitting estimation of standard errors for this impact coefficient. From equation (1d), we see that village dummies have been added to the specification. In their absence, correlation in the village-cluster-induced composite error will induce inefficiency, requiring an adjustment either by clustered standard errors, or by introducing village fixed effects. I follow the latter procedure, while simultaneously clustering standard errors at least at the region/state level. Although one can cluster along the unit of the village, this can be avoided by adding the village dummies. Standard error estimation in a treatment effect problem, and the appropriate unit of clustering is an active research area and always warrants further work.

At this stage, I introduce a relatively new econometric caveat that is not usually identified in empirical applications of this nature. A potential problem can emerge from the fact that the treatment or exposure variation is focused on a few household clusters or EAs in a section of the WR that closely surround the spill river. It is possible that due to this, clustered standard errors at any level of spatial aggregation may be misleading. To what extent this is really true in our case can only be determined by Monte Carlo analysis to examine the actual size of the test for the hypothesis relating to the shock impact coefficient. Thus, as future work for this paper, I designate the implementation of Conley-Taber standard errors (Conley and Taber (2009)) as yet another methodological extension to further improve upon the detection of the shock impact for this spill.

5 Quantile Regression Specification

5.1 Theory

In order to examine the treatment effect of the 2001 shock across the height distribution, while allowing for state and cohort fixed effects, and relevant control variables, I implement the model from Koenker (2010). This is an additive nonparametric model for quantile regres-

sion using the Laplacian fidelity instead of the Gaussian likelihood, and L_1-norm measuring total variation. The flexible structure not only allows for exploring treatment effects across the height distribution, but also control for additional risk factors, using both continuous and linear effects. Further, the regularized structure of the objective function enables me to impose a total variation penalty on the smooth covariates, and allows for embedding a LASSO penalty for the overall problem. Thus, the following conditional quantile model form is specified:

$$Q_{Y_i|x_i,z_i}(\tau|x_i,z_i) = x_i'\beta + \sum_{j=1}^{J} g_j(z_{ij}). \tag{4}$$

Here, the z_i's represent the nonlinear covariates. The nonparametric components g_j are assumed to be continuous functions, usually univariate, also possibly bivariate. Call the vector of these functions $g = (g_1, \ldots, g_J)$. Then, we can estimate g and $\beta \in \mathbb{R}^K$ solving:

$$\min_{\beta,g} \sum \rho_\tau(y_i - x_i'\beta - \sum g_j(z_{ij})) + \lambda_0 \sum_{k=1}^{K} |\beta_k| + \sum_{j=1}^{J} \lambda_j \bigvee(\nabla g_j) \tag{5}$$

where the first component is the quantile loss function, indexed by ρ which specifies the quantile at which we minimize the objective function; λ_0 is the LASSO tuning parameter allowing model selection for the parametric coefficients; and λ_j's are the tuning parameters for the nonparametric components. $\bigvee(\nabla g_j)$ is the total variation of the derivative on the gradient of g, and $\bigvee(\nabla g_j) = \| \nabla^2 g(z) \| dz$, where $\nabla^2 g(z)$ is the Hessian of g, and $\| \cdot \|$ denotes the Hilber-Schmidt norm for this matrix. The smooth components are subject to this penalty. In the univariate setting, these total variation penalties were suggested by Koenker et al. (1994) as an expedient smoothing device for nonparametric effects in quantile regression. In terms of notation, J is the number of nonlinear covariates, whereas K is the number of linear effects. The implementation of (5) is done using Koenker's `rqss` in `R` from Koenker et al. (1994) and Koenker and P. (2003).

5.2 Exposure Identification

The response variable used here is child height. The exposure covariate is included in the linear effects. Its spatial definition is more general, including those born in the WR. Relevant control factors like those specified and used in section 4, are also included. The following are expressed as the nonparametric smooth components: mother's age and education, and cohort. These can potentially extend a nonlinear effect, and modeling them smoothly allows for more flexible estimation of the treatment effect across quantiles. Finally, for incorporating GPS information uniquely, I introduce smooth effects of the latitude and longitude of villages in the sample. The embedded LASSO allows for a second level of model selection, with which we closely examine quantile treatment effects.

5.3 λ-selection and Results

With specification (5), as employed in Koenker (2010), the LASSO procedure outlines a regularization path, indexed by tuning parameter λ_0. At each point on the path, predictor coefficients are either driven to 0, or shrunk in value. Tuning parameter selection for the LASSO λ is done following Koenker's procedure of using a Schwartz-like criterion:

$$SIC(\lambda) = n\log\hat{\sigma}(\lambda) + \frac{1}{2}df(\lambda)\log(n) \tag{6}$$

where $\hat{\sigma}(\lambda) = \frac{1}{n}\rho_\tau(y_i - \hat{g}(x,z))$, and $\hat{g}(x,z) = x'\hat{\beta} + \sum_{j=1}^{J}\hat{g}_j(z)$. Further, as elucidated in Koenker (2010), $df(\lambda)$ is simply the effective model dimension, given by: $df(\lambda) = div(\hat{g}) = \sum_{i=1}^{n}\frac{\partial\hat{g}(x_i,z_i)}{\partial y_i}$. The quantity $df(\lambda)$ is defined as the trce of a pseudo-projection matrix in the

case of a linear least-squares-type model. Koenker also defines such a matrix for model (5):

$$\tilde{X} = \begin{pmatrix} X_0 & X_1 & \dots & X_J \\ \lambda_0 H_K & 0 & \dots & 0 \\ 0 & \lambda_1 P_1 & \dots & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_J P_J \end{pmatrix} \tag{7}$$

where X_0 is the matrix with the parametric linear covariates, X_j's are the basis expansions for the g_j's, $H_k = [0 \vdots I_K]$ consists of the contribution from the LASSO penalty, not including the intercept, and the P_j terms denote the contributions from the penalty terms on the smoothed components.

Since $\lambda \in \mathbb{R}^6$, where the first 5 values are smoothing parameters for the nonparametric terms, mother's age, education, cohort, latitude and longitude, and the last one is the LASSO tuning parameter, optimizing SIC is not trivial. A prudent strategy suggested by Koenker (2010) is to explore informally, narrow the region of optimization, and then follow a global optimization routine. In this paper, SIC optimization is implemented for the smoothing parameter values individually, and using these, the value for the LASSO λ is optimized. It is a one-dimensional procedure, although in forthcoming work, there will be multi-dimensional optimization for all the tuning parameters simultaneously, as seen Koenker (2010). I find the 6 optimal λ values to be $\{18, 7, 18, 6, 6, 13\}$.

I standardize the linear components of the design matrix. The findings are summarized in figures 8 and 9. Figure 8 plots the exposure coefficient for the deciles in $\tau \in [0.1, 0.9]$. It shows how including GPS information through smoothed latitude and longitude effects yields a more negative coefficient at almost every quantile. This is analogous to the finding in section 4 and Table 5, where a more accurate spatial definition of the exposure variable captures the shock impact (column (7)). Figure 9 plots the exposure coefficient, with standard error intervals, for this final quantile treatment model with smoothed GPS effects, for the percentiles in $\tau \in [0.1, 0.9]$.

6 Conclusion

In this paper, I use 2003 DHS data on Ghana, to examine the health consequences of environmental shocks in the form of chemical spills from mining activity in WR in 2001, on children born during the shock. The paper contributes to shock impact evaluation literature by combining survey data with the associated GPS data, to better spatially and temporally identify exposed children, especially since the survey data is at a higher level of aggregation. In addition to specifying spatial exposure in fixed effects models, this paper introduces 2 other unique ways of incorporating this GPS information: first, examining child height as a bi-variate function smoothed location coordinates to reflect this relationship in a contour map; second, incorporating these smooth components in a flexible additive quantile regression model to investigate exposure impact across the whole height distribution, instead of only at the mean. I find that at all levels, and across all empirical strategies, embedding GPS information creates a stronger link between shock variables and health outcomes. This has beneficial methodological implications, by way of improved procedures for modeling shock-health outcome link in the environmental health and epidemiology literature. It is especially relevant to studying smaller-scale shocks which need a localized specification of the treatment.

A new econometric caveat introduced here is the need for a continued re-examination of the standard errors, acknowledging the fact that exposure variation is focused on fewer groups relative to the number of control groups. Applications of this nature may not focus on this issue. As yet another extension of following procedures that will derive an improved examination of the shock impact of this spill, I designate the implementation of the Conley-Taber standard errors as future work related to this paper.

The analysis in this paper demonstrates the importance of linking environmental health problems related to livelihood-based shocks, with the larger issue of following the right methodological path to studying these questions. Through this process, there is a large potential of better informing policy decisions at all levels of governance. Originally motivated by the attractive theoretical properties of the procedures described above, the primary con-

tribution of this paper is in an expansion of the mechanics related to estimating the impact of a shock on health.

References

Akaike, H. (1974), 'A new look at the statistical model identification.', *IEEE transactions on Automatic Control* **AC-19**, 716–723.

Akresh, R., P. Verwimp and T. Bundervoet (2007), 'Civil war, crop failure, and child stunting in Rwanda', *Economic Development and Cultural Change (Forthcoming)* .

Alderman, H. and J. Behrman (2006), 'Reducing the incidence of low birth weight in low income countries has substantial economic benefits', *World Bank Research Observer* **21**(1), 25–48.

Alderman, H., J. Hoddinott and B. Kinsey (2006), 'Long term consequences of early childhood malnutrition', *Oxford Economic Papers* **58**(3), 450–74.

Almond, D., L. Edlund, Li. Hongbin and J. Zhang (2007), Long-term effects of the 1959-1961 China Famine: Mainland China and Hong Kong, NBER Working Paper 13384, National Bureau of Economic Research, Cambridge MA.

Bertrand, M., E. Duflo and S. Mullainathan (2004), 'How much should we trust differences-in-differences estimates?', *Quarterly Journal of Economics* **119**(9), 249–275.

Boadi, N.O., S.K. Twumasi and J.H. Ephraim (2009), 'Impact of cyanide utilization in mining on the environment', *International Journal of Environmental Resources* **3**(1), 101–108.

Bundervoet, T., P. Verwimp and R. Akresh (2009), 'Health and civil war in rural Burundi', *Journal of Human Resources* **44**(2), 536–563.

Conley, T.G. and C. Taber (2009), 'Inference with 'Difference in Differences' with a small number of policy changes', *NBER Working Paper* .

Duflo, E., M. Greenstone and R. Hanna (2008), 'Indoor air pollution, health and economic

well-being', *SAPI EN. S. Surveys and Perspectives Integrating Environment and Society* (1.1).

Efron, B., T. Hastie, I. Johnstone and R. Tibshirani (2004), 'Least angle regression', *Annals of Statistics* **32**(2), 407–451.

Fan, J. and Li. R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of American Statistical Association* **96**, 1348–1360.

Godfrey, Keith M. and Barker David J.P. (2000), 'Fetal nutrition and adult disease', *American Journal of Clinical Nutrition* **71**(5), 1344S–1352S.

Huang, J., J.L. Horowitz and S.G. Ma (2008), 'Asymptotic properties of bridge estimators in sparse high-dimensional regression models', *Annals of Statistics* **36**, 587–613.

Kézdi, Gábor (2004), 'Robust standard error estimation in fixed-effects panel models', *Hungarian Statistical Review* **9**, 96–116.

Koenker, R. (2010), *Advances in Social Science Research Using R*, Lecture Notes in Statistics 196, Springer Science+Business Media, LLC, chapter Additive Models for Quantile Regression: An Analysis of Risk Factors for Malnutrition in India.

Koenker, R. and Ng P. (2003), 'SparseM: A Sparse Linear Algebra Package for R', *J. Stat. Software* **8**(6), 1–9.

Koenker, R., P. Ng and S. Portnoy (1994), 'Quantile Smoothing Splines', *Biometrika* **81**, 673–680.

Lin, Y. and H. Zhang (2006), 'Component selection and smoothing in multivariate nonparametric regression', *Annals of Statistics* **34**, 2272–2297.

Maccini, S. and D. Yang (2008), 'Under the weather: Health, schooling and economic consequences of early-life rainfall', *American Economic Review* .

Martorell, R. and J. Habicht (1986), *Growth in Early Childhood in Developing Countries*, Plenum Press.

Meier, L., van der Geer S. and P. Bülmann (2006), 'The group lasso for logistic regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 53–7.

of Ghana, Government (2008), Implementation of the Ghana Poverty Reduction Strategy, Annual progress report, National Development Planning Commission.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Tschakert, P. and K. Singha (2007), 'Contaminated identities: Mercury and marginalization in Ghana's artisanal mining sector', *Geoforum* **38**(6), 1304–1321.

Van den Berg, G., M. Lindeboom and F. Portrait (2006), 'Economic conditions early in life and individual mortality', *American Economic Review* **96**(1), 290–302.

Vieira, V.M., J.E. Hart, T.F. Webster, J. Weinberg, R. Puett, F. Laden, K.H. Costenbader and E.W. Karlson (2010), 'Association between Residences in US Northern Latitudes and Rheumatoid Arthritis: A Spatial Analysis of the Nurses' Health Study', *Environmental health perspectives* **118**(7), 957.

Yuan, M. and Y. Lin (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67.

Zhao, P., G. Rocha and Y. Bin (2009), 'The composite absolute penalties family for grouped and hierarchical variable selection', *The Annals of Statistics* **37**(6A), 3468–3497.

Zou, H., T. Hastie and R. Tibshirani (2007), 'On the "degrees of freedom" of the lasso', *Annals of Statistics* **35**(5), 2173–2192.

Figure 1: Map of Ghana.

Figure 2: Map of Western Region, Ghana.
Source: http://www.nationsonline.org.

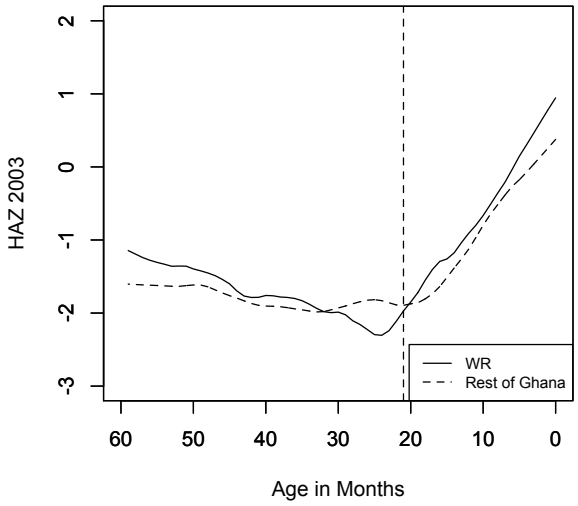


Figure 3: Height-for-Age Z scores by shock region (WR) and birth cohort.

Notes: Nonparametric locally weighted regression of height-for-age z scores on child age, for children born between 1998 and 2003, comparing those born in WR with those born in the rest of Ghana. The vertical dashed line indicates the end of 2001, the spill year.

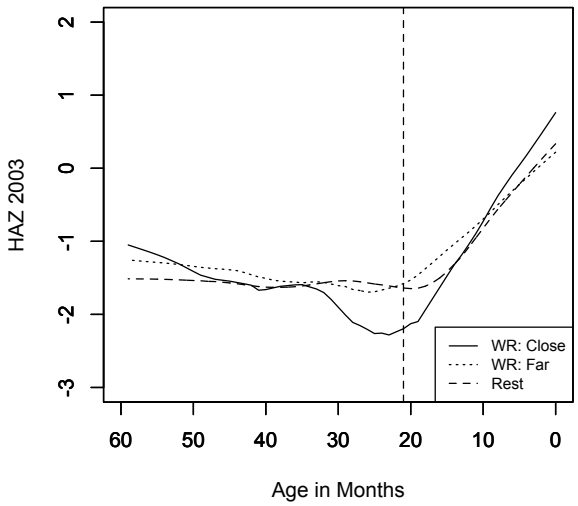


Figure 4: Height-for-Age Z scores by EAs in WR (close/far to spill) and birth cohort.

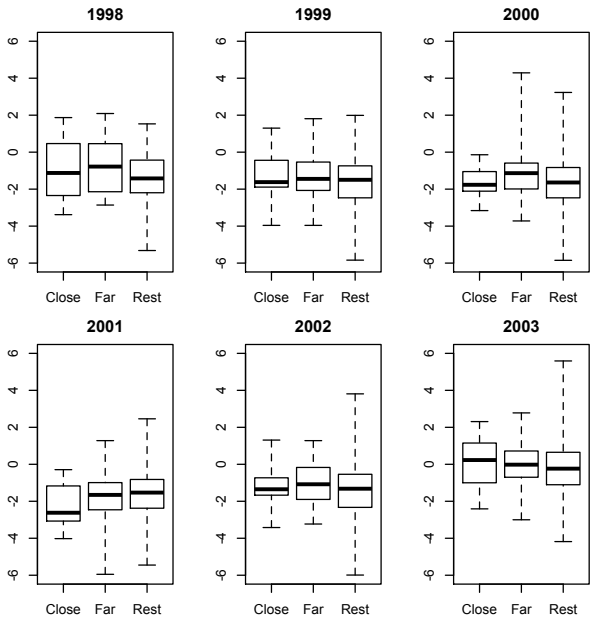


Figure 5: Height-for-Age Z scores by shock region (WR) and birth cohort: Box-plots.

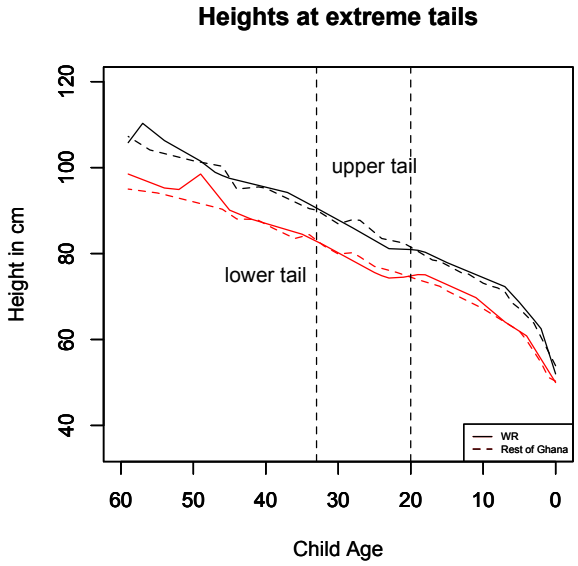
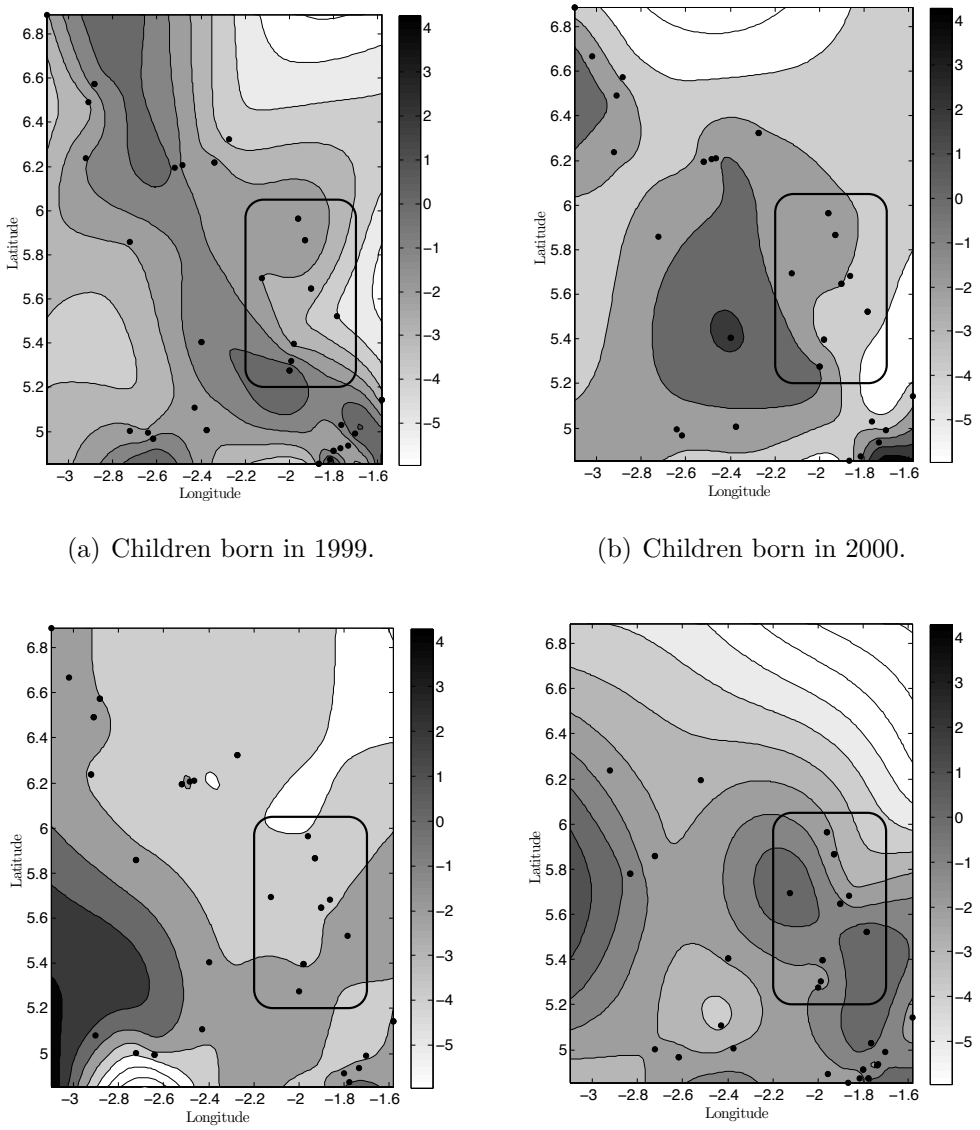


Figure 6: Height in cm by shock region and birth cohort: extreme tails.



(a) Children born in 1999.

(b) Children born in 2000.

(c) Children born in 2001.

(d) Children born in 2002.

Figure 7: Contour map of height-for-ages for children across Western Region.

Notes: Nonparametric (multivariate) regression of height-for-age Z scores on the GPS coordinates of the child's birth household. The spill occurred slightly upstream from the villages with coordinates (5.374, -2.016) and (5.382, -2.063). The major mining town Tarkwa, is located at (5.3,-1.983). The contour colors indicate the level of the height-for-age for the clusters included in those contours. The legend links these colors to the height-for-age level. Height-for-age begin appearing in worse (lighter) contours for all children born during or before 2001. The rectangle encloses the HH clusters closer to the spill site, and thus *exposed*. Black dots are actual HH clusters.

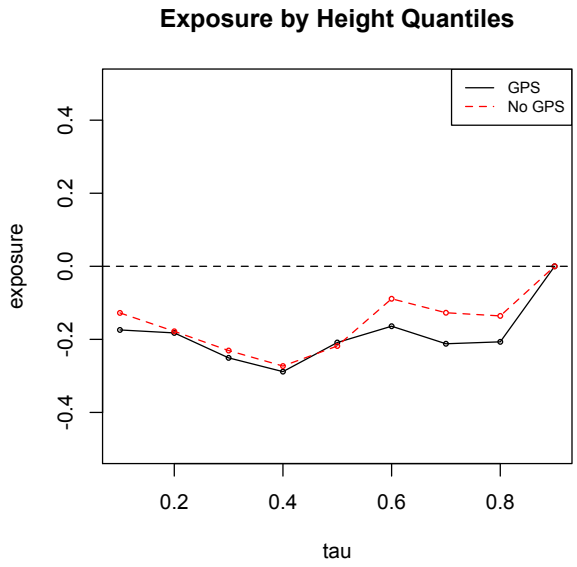


Figure 8: Exposure coefficient for deciles: $\tau \in [0.1, 0.9]$.

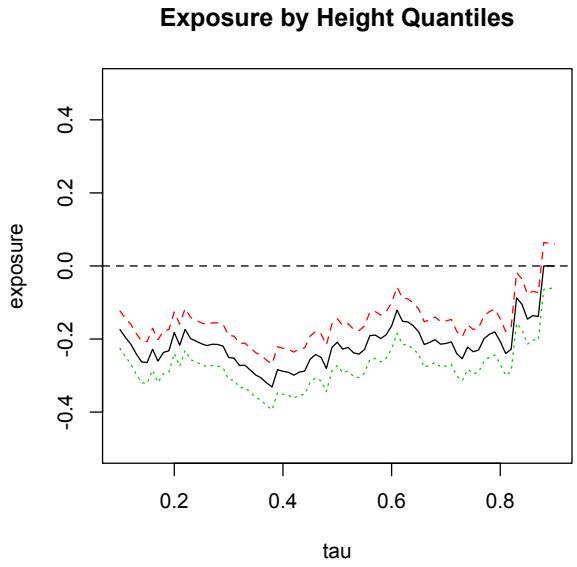


Figure 9: Exposure coefficient with confidence intervals, for percentiles: $\tau \in [0.1, 0.9]$.

Table 1: Comparing key criteria across states.

	% RWS		%Poor		%Low BW		% Stunting Prevalence			
							Urban		Rural	
States	2005	2008	2003	2008	2003	2008	2003	2008	2003	2008
Western	43.7	41.3	47.8	48.3	46.8	57.1	11.1	31.1	41.4	23.7
Central	36.5	44.3	50.8	36.9	16.6	50.0	21.7	28.1	38.3	31.9
Greater Accra	59.0	65.0	8.3	6.7	45.3	42.3	12.6	15.9	28.5	23.8
Volta	61.3	54.3	54.3	57.1	59.3	44.4	26.6	11.6	33.8	27.8
Eastern	53.2	58.9	50.8	46.6	33.9	35.6	19.0	27.9	37.2	36.0
Ashanti	50.2	72.9	36.8	37.4	47.3	56.2	19.3	19.8	34.9	36.7
Brong Ahafo	46.6	53.5	55.6	58.4	41.3	53.4	33.8	14.7	41.1	30.6
Northern	47.9	57.9	80.7	78.9	28.3	36.5	40.7	25.6	57.5	32.7
Upper East	69.0	52.2	88.8	84.1	42.2	50.0	46.6	38.9	38.7	42.3
Upper West	93.1	76.8	84.2	81.4	9.3	43.7	44.4	15.3	36.9	29.2

% RWS: Rural Water Supply.
% Low BW: Proportion of Low Birth-Weight babies.

Table 2: Height-for-age Z-scores across states: 2003 DHS.

	n	Min	10%	25%	$Median$	$Mean$	75%	90%	Max
Western	283	-5.95	-2.97	-2.13	-1.27	-1.21	-0.32	0.77	4.29
Central	204	-5.60	-2.96	-2.18	-1.23	-1.33	-0.42	0.23	2.91
Greater Accra	248	-4.88	-2.43	-1.53	-0.71	-0.65	0.17	1.20	5.59
Volta	204	-4.17	-2.82	-1.98	-1.10	-1.11	-0.19	0.59	3.04
Eastern	254	-5.99	-2.73	-2.01	-1.21	-1.16	-0.32	0.47	5.11
Ashanti	451	-5.51	-3.03	-2.12	-1.28	-1.29	-0.41	0.47	3.23
Brong Ahafo	352	-5.85	-3.07	-2.25	-1.27	-1.32	-0.39	0.42	2.16
Northern	450	-5.84	-3.60	-2.89	-1.94	-1.93	-1.03	-0.30	2.99
Upper East	269	-5.20	-3.11	-2.22	-1.35	-1.27	-0.45	0.49	4.13
Upper West	191	-4.55	-2.93	-2.10	-1.24	-1.19	-0.36	0.68	3.17
Ghana	2906	-5.99	-3.07	-2.23	-1.31	-1.30	-0.40	0.49	5.59

Table 3: Height-for-Age Z scores: By Region and 2001 Shock Exposure.

A. All Children	Shock Region $n = 283$	Rest of Ghana $n = 2623$	Difference
Not Born During Shock	-1.0676	-1.3577	0.2880***
	(0.0101)	(0.0430)	(0.1347)
Born During Shock	-2.2693	-1.9313	-0.3380
	(0.0234)	(0.0028)	(0.1884)
Difference	-1.2017**	-0.5756***	-0.6260***
	(0.2056)	(0.0715)	(0.2177)
B. Girls	Shock Region $n = 140$	Rest of Ghana $n = 1429$	Difference
Not Born During Shock	-0.8624	-1.1825	0.3201**
	(0.0137)	(0.0489)	(0.1497)
Born During Shock	-1.9439**	-1.8577**	-0.0862
	(0.0372)	(0.0052)	(0.2293)
Difference	-1.0815	-0.6752***	-0.4062
	(0.2565)	(0.0959)	(0.2739)
C. Boys	Shock Region $n = 143$	Rest of Ghana $n = 1477$	Difference
Not Born During Shock	-1.2503	-1.5210	0.2706*
	(0.0136)	(0.0483)	(0.1563)
Born During Shock	-2.7303	-2.0061	-0.7242**
	(0.0578)	(0.0059)	(0.2932)
Difference	-1.4800**	-0.4851***	-0.9947***
	(0.3149)	(0.1059)	(0.3323)

Notes: Robust clustered standard errors. * significant at 10%, ** significant at 5%, *** significant at 1%. The shock region includes the whole Western Region (WR), Ghana, where the chemical spill occurred in October 2001, near Tarkwa. Rest of Ghana excludes the WR. The model considers only those born in 2001 as exposed. This is in contrast with other successive specifications that consider *all* children born before and alive through October 2001, and those located in clusters closer to the spill as exposed, and thus better identify shock exposure.

Table 4: Region-Birth Cohort Fixed Effects: Impact of the 2001 Shock on Children's HAZ Scores.

	Dependent Variable: Children's HAZ scores					
	(1)	(2)	(3)	(4)	(5)	(6)
Exposed	-0.510***	-0.365**	-0.593***	-0.357***	-0.239*	-0.280*
	(0.067)	(0.147)	(0.085)	(0.057)	(0.109)	(0.149)
Male	-0.203***	-0.154*	-0.196**	-0.197**	-0.203***	-0.203***
	(0.057)	(0.076)	(0.081)	(0.055)	(0.057)	(0.057)
Water	-0.155	-0.223**	-0.144	-0.211**	-0.206**	-0.210**
	(0.101)	(0.090)	(0.141)	(0.083)	(0.082)	(0.083)
Rural	-0.077	-0.181**	-0.229	-0.158*	-0.159*	-0.156*
	(0.073)	(0.071)	(0.135)	(0.074)	(0.074)	(0.074)
Male × Exposed				-0.309**		
				(0.070)		
Water × Exposed					-0.305**	
					(0.110)	
Rural × Exposed						-0.253*
						(0.131)
Bad Cooking Fuel	-0.254***	-0.247***	-1.276***	-0.291***	-0.289***	-0.290***
	(0.047)	(0.071)	(0.273)	(0.054)	(0.053)	(0.053)
Region Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Cohort Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
n	2670	1158	1982	2670	2670	2670

Notes: Robust standard errors, clustered at region level. Significant at * 10%, ** 5%, *** 1%.
Water: Indicator for ground/surface water being primary household water source.
Exposed = WR × Born During Shock.
Columns (2) and (3) estimate the baseline model (column (1)) for rich and poor households, respectively. Column (4) - (6) add exposure intensity measures: being a male child, exposed to natural water sources, and being a rural resident, respectively.

Table 5: Region/Cluster-Birth Cohort Fixed Effects: Impact of the 2001 Shock on Children's HAZ Scores.

	Dependent Variable: Children's HAZ scores						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Exposed	-0.074	0.040	0.472***	-0.188***	-0.211**	-0.157**	-0.163**
	(0.063)	(0.058)	(0.072)	(0.048)	(0.060)	(0.050)	(0.062)
Male	-0.201***	-0.187***	-0.200***	-0.201***	-0.233***	0.197**	-0.190
	(0.051)	(0.050)	(0.050)	(0.051)	(0.051)	(0.060)	(0.059)
Water	-0.428***	-0.427***	-0.371***	-0.426*	-0.218	-0.193**	-0.183
	(0.094)	(0.093)	(0.071)	(0.093)	(0.174)	(0.097)	(0.167)
Male × Exposed		-0.223***					
		(0.065)					
Water × Exposed			-0.736***				
			(0.074)				
Bad Cooking Fuel	-0.524***	-0.529***	-0.504***	-0.526***	-0.172*	-0.149	0.018
	(0.055)	(0.052)	(0.063)	(0.053)	(0.079)	(0.084)	(0.153)
Region Fixed Effects	Yes	Yes	Yes	Yes	No	Yes	No
Cluster Fixed Effects	No	No	No	No	Yes	No	Yes
Cohort Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
n	2670	2670	2670	2670	2670	2670	2670

Notes: Robust standard errors, clustered at region level. Significant at * 10%, ** 5%, *** 1%. *Water*: Indicator for ground/surface water being primary household water source. All children in the survey who are born before the shock, and are alive through it, are considered as exposed. There are 3 forms of spatial and specification-based model groupings in this table:

(a) Grouping 1: Columns (1) - (3), *Exposure = WR × Alive through Spill* (where columns (2) and (3) add interactions measuring intensity of exposure of shock through gender and water source type). Columns (4) - (7), *Exposure = Closer to Spill × Alive through Spill*.

(b) Grouping 2: Columns (1) - (4) and (6) have region fixed effects, and columns (5) and (7) have village fixed effects. This leads to: control of within-cluster variation, a higher treatment effect.

(c) Grouping 3: Columns (1) - (5) contain insufficient additional control variables. Columns (6) and (7) add a rich set of controls, based on maternal and socioeconomic factors. This leads to: better resolution of omitted variable bias, and no difference between the treatment effect in columns (6) and (7).