Measuring Flood Resilience using Remote Sensors

Víctor Funes-Leal

Abstract

What are the effects of floods on reporting likelihood and observable outcomes? I examine this question in the context of a Randomized Control Trial (Shukla and Baylis, 2019) aimed at adopting a specific new technology for small-scale farmers in Bihar, India. I study two effects; first, to which extent adaptation to a regular rainfall pattern (the South Asian Monsoon) makes farmers under-report the impact of floods/heavy rainfall. To do so, I use inundation maps from satellite-measured floodwater to compare observed and reported floods. Second, given that I can determine which household lives near flooded areas, I measure their impact on food security outcomes. On the one hand, there is significant evidence in favor of under-reporting bias, but I also find little evidence of impacts on food security outcomes.

Keywords: Floods, Satellite Data, Reference Dependence, Food Security

JEL Codes: Q54, O18, D83, Q15

Introduction

How does climate change affect poor households' welfare? A large share of Developing Countries' population is usually engaged in small-scale farming, so extreme weather events such as heatwaves, floods, and droughts affect them significantly.

Poor farmers in flood-prone areas are likely to experience large adverse shocks because their geographical location is a function of land prices, which are likely to decrease as a consequence of floods; farmer's livelihood depends on their production, which is a nonlinear function of yields, a larger-than-usual rainfall will harm them and will also generate more considerable post-harvest losses.

Floods have other negative impacts besides lower income due to decreased yields; they may reduce farmer's capital (increased cattle mortality, damage to buildings and tools). Consequently, weather shocks will reduce farmer's food security, and the lack of well-functioning insurance markets exacerbates this effect.

I can identify several market failures that affect these households: first, anthropogenic climate change, which is an externality at a global scale but has heterogeneous effects within and between countries (poor households in developing countries are more likely to experience a significant reduction in their welfare relative to wealthier households in their countries and elsewhere). Second, several issues related to asymmetric information can exacerbate the negative impact I have described; for example, the lack of a well-functioning system of weather alerts to warn farmers about higher-than-usual rainfall prevents them from taking preemptive measures to cope with floods.

Similarly, as pointed out by Spence et al. (2011), individuals who have experienced floods in the past have different behavior patterns than those who have not, the former showing a higher degree of risk-aversion and perceived vulnerability to their effects. However, Guiteras et al. (2015) argues that self-reported flood experience is subject to recall error and reference dependence, among other biases, thus leading to under-reporting of flood damage.

Kocornik-Mina et al. (2020), using panel data spanning most of the world over 35 years, show that floods have only a short-term impact on economic activity in cities; after an extreme weather event, low-elevation areas recover as fast as their high-elevation counterparts. Unlike theirs, my sample consists of farmers living in small villages; hence, it can be argued that their results cannot be extrapolated linearly to small villages and rural areas; in other words, are farmers exposed to the same "equalizing" forces that lead urban areas to recover from floods?

This article provides evidence for two phenomena: first, farmers in Bihar tend to underreport the impact of floods because flood presence hurts flood reports. Second, floods have detrimental effects on food security. Still, they are curbed by farmers' avoidance behavior, which is a consequence of adaptation to a weather pattern that consists of a large concentration of rainfall in a short period (the South Asian Monsoon).

Experimental Design

Details

As detailed in Shukla and Baylis (2019), their intervention aimed to explore the extent to which user experience and reference frame effects impact willingness to pay for a productivity-enhancing new technology. The new technology in question was hermetic storage bags which allow farmers to store grains for more extended periods but cost almost seven times more than the storage technology used until then (jute bags). Hermetic bags were allocated using a two-stage randomized price experiment with a Becker-DeGroot-Marschak (BDM) auction, allowing the authors to separate the effects of experiential learning and reference dependence.

The Randomized Control Trial, as described by Shukla and Baylis (2019) and Shukla et al. (2023), was carried out in the Indian State of Bihar in three different rounds from November 2015 (baseline round) to January 2019 (endline round) also including a midline round in July 2017. The area encompassed 80 villages from five districts (Purba Champaran, Samastipur, Begusarai, Bhagalpur, and Banka); from these, 42 villages were selected for the auction, for a total of 1429 households, for the remaining 38 villages (2571 households), the treatment was assigned at random.

Figure 1
Spatial distribution of households

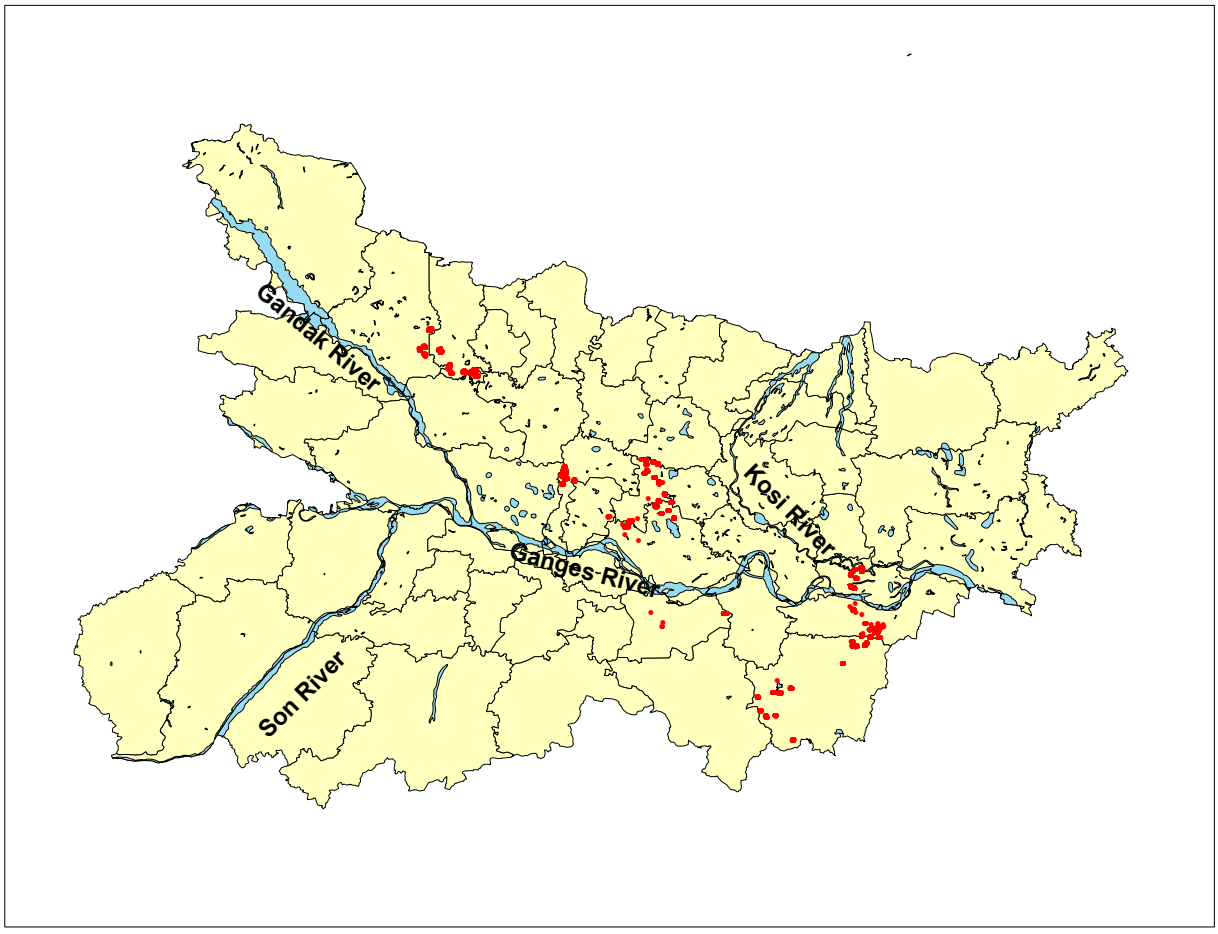
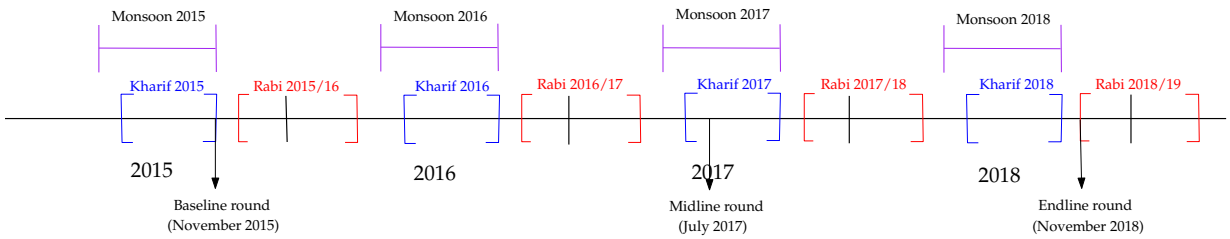


Figure 1 shows the coordinates of every household in the sample; all households were georeferenced when the survey rounds were carried out, and the village-level clustering is evident. Unfortunately, I do not have data on the specific location of their farms; due to this fact, I will have to assume that farms are located in the vicinity of households.

The survey timeline (Figure 2) was:

- Baseline round: November 2015 - November 2016.
- Midline round: July 2017.
- Endline round: November 2018 – January 2019.

Figure 2
Survey timeline



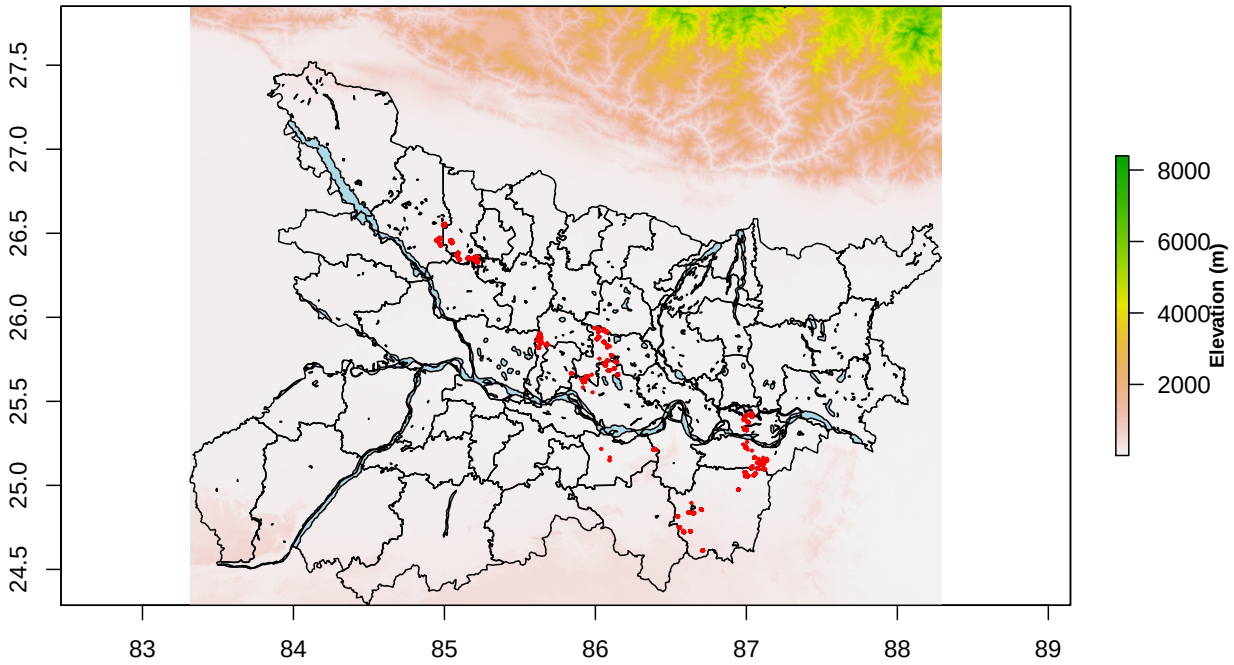
Farmers in Bihar grow wheat, maize, and rice; on the baseline survey 92% of farmers in the sample either rented the land or owned less than a hectare of land, their average number of plots was 1.9, 44% of household heads were illiterate, and the average family size was 4.1 persons.

Given the small scale of their farms, households do not have access to storage technology such as silos. Instead, they store grains in jute bags to sell their harvest. These jute bags are prone to fungal growth and pest infestation, thus leading farmers to sell their grains at discounted prices and reducing their incomes to values below those they would have otherwise. Fungal growth also leads to problems related to aflatoxin contamination; chronic exposure to them leads to compromised immunity: liver, kidney, and spleen enlargement; congenital disabilities; and several carcinogenic effects.

Data

Elevation estimation

Figure 3
Topographical map of Bihar (SRTM data)



Besides the geographical coordinates of households, the survey also contains their estimated orthometric height measured by the GPS device. However, later analyses showed that these measurements were faulty because their height readings differed significantly between midline and endline rounds due to differences in atmospheric conditions.

A solution to this problem is to use interpolated height data from a digital elevation model (Farr et al., 2007), provided by the Shuttle Radar Topography Mission[1], which contains elevation data from 56°S to 60°N latitude with a 90 meters resolution, which is suitable for height estimation. SRTM uses a synthetic aperture radar interferometer to create a digital elevation model on a satellite. Height is measured based on the discrepancy between this data and the WGS84 ellipsoid.

The Consortium for Spatial Information (CGIAR)[2] released processed data from the SRTM as raster files covering a $5^\circ \times 5^\circ$ tile, since Bihar is located in the intersection of four tiles, the data plotted in Figure 3 was obtained by combining them into a "mosaic" and then cropping the file up to the required extent.

[1] https://www2.jpl.nasa.gov/srtm
[2] http://srtm.csi.cgiar.org/srtmdata

Estimated height per household is obtained by a bilinear interpolation from four contiguous pixels since height does not vary significantly across the state due to its location within the Indo-Gangetic plains; similarly, a buffer can be defined around each point, but for the reason above, its result will not be significantly different.

Rainfall variability

The Indo-Gangetic Plain is an alluvial plain where monsoon rains along the Ganges basin lead to flooding. For example, the 2017 monsoon season produced many flash floods across northern districts, accounting for about 500 casualties and significant damages to private property and public infrastructure such as railroads, roads, and bridges.

One way to measure rainfall variations is using data from the **Climate Hazards Center InfraRed Precipitation with Station data** (CHIRPS) (Funk et al., 2015), which is a quasi-global rainfall data set from over 30 years. Spanning 50°S-50°N (and all longitudes), starting in 1981 to the present, CHIRPS incorporates 0.05° resolution satellite imagery with in-situ station data to create gridded rainfall time series for trend analysis and seasonal drought monitoring. This data set consists of daily raster files that were first clipped to include only pixels inside the Bihar shapefile[3] and aggregated by month/year and district.

Rainfall variability across districts is also an issue for the analysis because farmers are distributed across several districts. To account for this variation, I adopted the method described by Bandyopadhyay and Skoufias (2015), which consists of measuring rainfall variability using the coefficient of variation (standard deviation over mean) across the entire period and then defining year t as a flood (drought) year if rainfall in t is higher (lower) than the 80th (20th) percentile. A second option is the one from Marchetta et al. (2019), which consists of calculating standardized rainfall deviations, defined as the difference between total year rainfall and its long-term mean, divided by its standard deviation:

$$RV_{dt} = \frac{RAIN_{dt} - \overline{RAIN}_d}{SD(RAIN)_d}$$

After this calculation, I can define a drought (flood) dummy equal to 1 if RV_{dt} is higher (lower) than the 80th (20th) percentile of the distribution. I used data from 1988 to 2018 to calculate both "long-term" means and standard deviations per district for the years 2015-2018.

Notice that both approaches use deviations from long-term means rather than levels; the rationale is the following: every farmer knows that the total rainfall during the monsoon season is significant enough to cause floods; it is already accounted for, unusually large deviations from its trend can be considered as unexpected and hence, exogenous to the farmer's decision process and thus can be used in an identification strategy.

[3] All shapefiles used for mapping state boundaries, rivers, and water sources come from the GADM database https://gadm.org.

Figure 4
Rainfall deviation index (RV_{dt}) by year and district

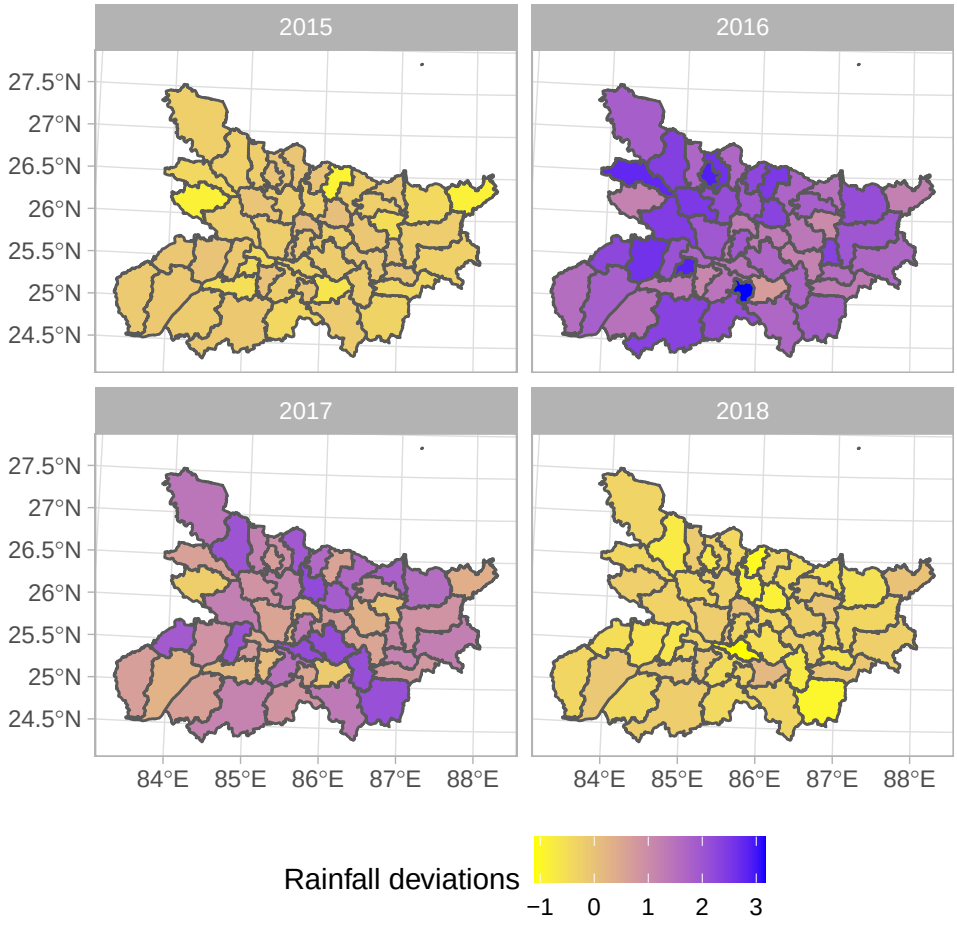
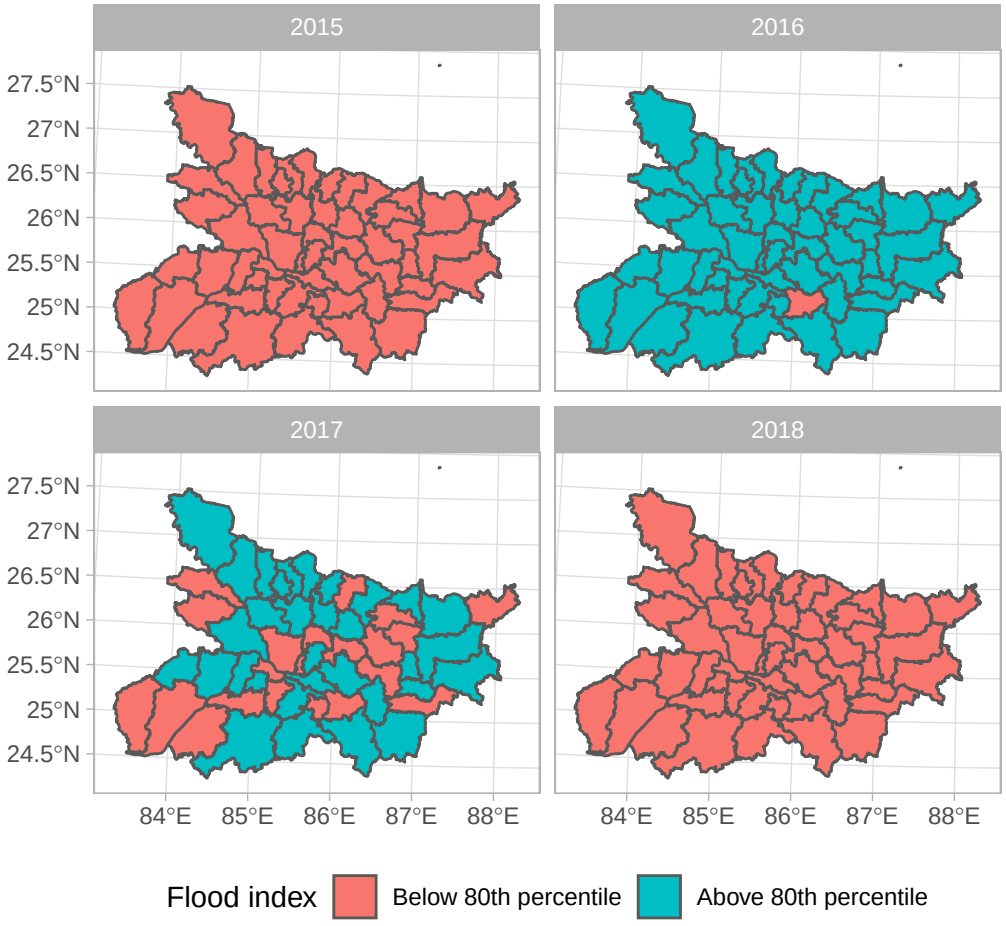


Figure 4 shows that 2016 and 2017 were relatively "wet" years compared to 2015 and 2018. On the former, almost all districts reported above-average normalized rainfall index, and in several cases, way above the average, as shown in Figure 5.

Figure 5
District with rainfall deviations above 80th percentile



Almost all districts experienced rainfall deviations on or above the 80th percentile in 2016, and most of them had similar results in 2017, with the opposite taking place in 2016 and 2018.

Flood detection using remote sensors

I need a reliable measure of flood extent; for that reason, I used data from remote sensors because they are freely available and exogenous to the survey mechanism. The use of satellite data is a new tool in Economics since its widespread availability is very recent (Donaldson and Storeygard, 2016), remote sensors give researchers access to consistent data across borders at a low or zero cost, but they have several issues to take into account when used to draw inferences, as detailed by Jain (2020), first, raw data must be preprocessed to correct for variations in sensor characteristics, this is performed through atmospheric corrections and radiometric calibrations by the data provider. Second, the data must be filtered to correct for cloud cover, which may lead to another kind of measurement error, and this is usually performed via cloud-removing algorithms. Finally, it is recommended to run validation analyses on the data to assess their accuracy; if that is not the case, the data may yield several "false positives" due to, for example, irrigated areas being classified as flooded.

I used data from the NASA Moderate Resolution Imaging Spectroradiometer - MODIS

(https://modis.gsfc.nasa.gov), an array of satellites that scan the earth's surface every two days recording reflectance values over 36 bands in the visible and infrared spectra. Data is then processed into cloud-free composites of 16 days at the 250m $\times$ 250m resolution. Two measures can be constructed from this set of time-indexed pixels, one that is sensitive to surface water and the other to surface vegetation, if the values of the first index are greater than those of the second, then the area is classified as having overlaying surface water.

This data from MODIS is processed by NASA's **MODIS Near Real-Time Global Flood Mapping Project** (NRT)[4] which releases global daily surface and flood water maps at the same 250m $\times$ 250m resolution as the original data as a raster (pixelated grid) where every pixel is classified as flooded or not, as stressed out by Jain (2020), pre-processing of satellite data to correct for cloud cover and haze is a key step to avoid misclassification issues.

The data collected ranges from January 2015 to December 2018, including four monsoon seasons; since my interest is to detect flood water, I used the MODIS Flood Water Product, which is defined as the total surface water detected minus the "Reference Water" which amounts for the off-season river extent.

MODIS generates data for 36 spectral bands at different resolutions (250 meters for bands 1-2, 500 meters for bands 3-7, and 1000 meters for bands 8-36); classification is performed by a ratio calculation using data from bands 1 and 2 in the following manner, let p_i ba any pixel i in a specified section of the planet, then area i is flooded if:

$$p_i = \frac{\rho_{2i} + \alpha}{\rho_{1i} + \beta} \geq \theta$$

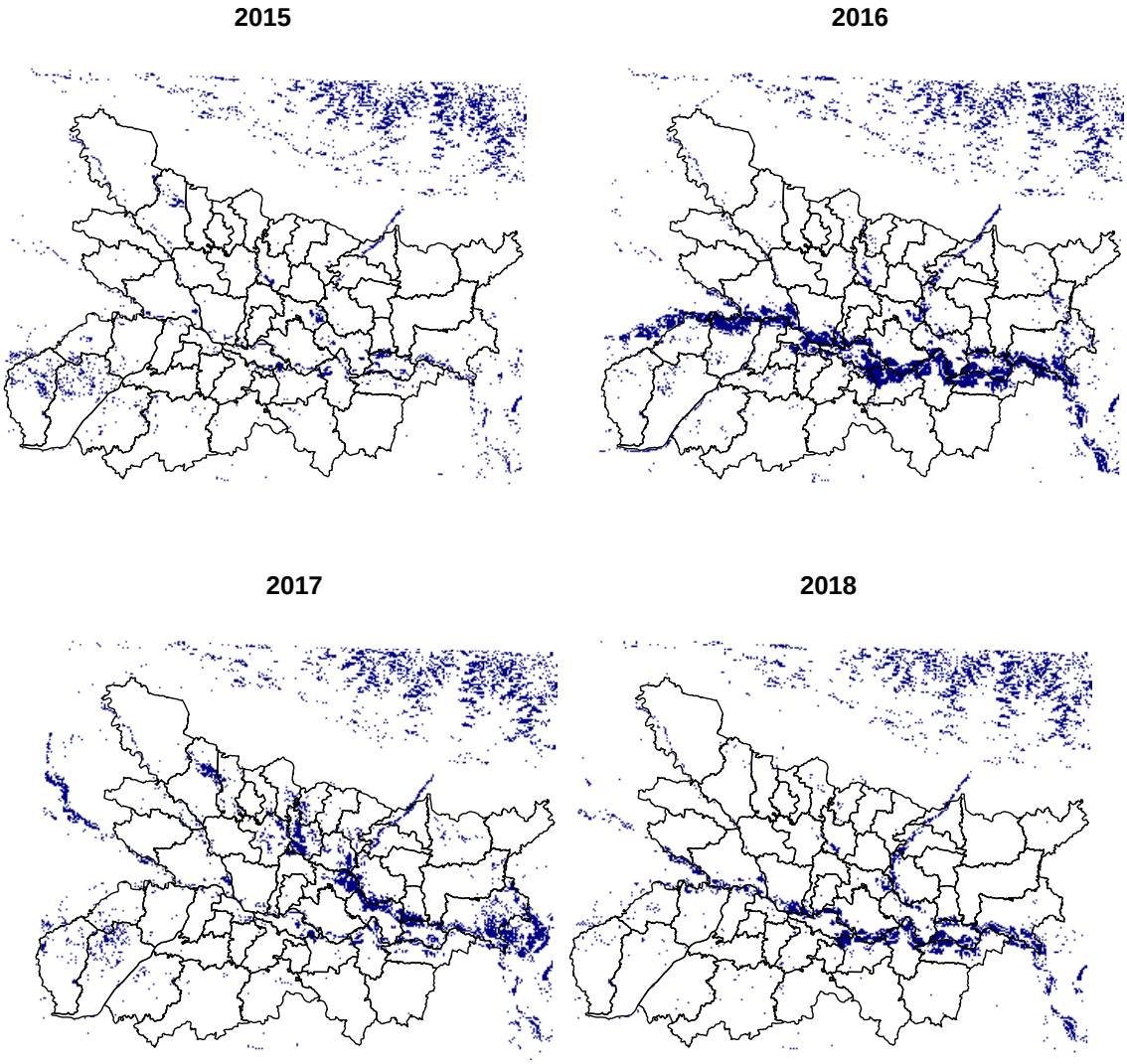
Where ρ_{1i} and ρ_{2i} are the reflectance at bands 1 and 2, θ is a threshold value varying from 0.6 to 0.9, α and β are constants determined empirically, other thresholds in bands 1 and 7 are used to correct for cloud cover[5].

MODIS produces several raster files with different temporal aggregation levels (one, three, and fourteen days). Since I want to minimize the impact of "patchiness" created by the cloud cover, I used the bi-weekly data for every month (24 files per year). Finally, I combined those 24 raster files into one single file per year (called a "raster stack"), which I used to aggregate by pixel across all dates and define a dummy variable equal to 1 if a specified pixel was flooded at least during one of the bi-weekly periods and zero otherwise.

[4] https://floodmap.gsfc.nasa.gov

[5] https://floodobservatory.colorado.edu/Tech.html

Figure 6
Flooded areas in Bihar (blue)



The full extent of satellite-detected floods for 2015-2018 is shown in Figure 6; this plot coincides with the findings on precipitation plots, 2016 and 2017 had more extensive precipitation indexes and, similarly, had more flooded areas. However, flood patterns are different, The 2016 floods were concentrated in the Ganges river basin, but the 2017 floods originated in the Kosi river basin.

Survey data

Food Security Outcomes

The survey contains data from household characteristics, food expenditure by categories, and crop sales. A module in the survey contains detailed information about food consumption, including the number of days each household consumed a series of food categories (rice, tubers, cereals, vegetables, fruits, lentils, eggs, dairy, meat, poultry, fish, sugar,

nuts, and packaged food), given this data, two indicators can be constructed for midline and endline rounds:

1. **Household Dietary Diversity Score (HDDS)**: this continuous score consists of the sum of all groups consumed Kennedy et al. (2010), Swindale and Bilinsky (2006), each category has a score equal to one if it was consumed by each household at least once during the reference period (last seven days before the survey date, this indicator requires a 24-hour recall, but as Jones et al. (2014) the survey uses a weekly recall period); hence this score is a number ranging from 0 (no food in those categories was consumed) to 15 (all categories were consumed at least once during the previous week).

   The HDDS is defined as "the ability to acquire sufficient quantity and quality of food to meet all household members' nutritional requirements for a productive life" Leroy et al. (2015); this indicator has a significant drawback: no cutoff point has been defined to determine which households have low/inadequate food diversity.

2. **Food Consumption Score (FCS)**: This score is equal to the frequency-weighted consumption of different food groups by a household during the seven days before the survey, this indicator contains information that the HDDS does not, first, it is weighted by the nutritional value of each category as a means to quantify both quantity and quality of household food access, since the weights were defined in terms of their "nutrient density" (in terms of total calories), thus higher weights imply higher caloric intake. Second, it has clearly defined cutoff points, the index ranges from 0 to 112 and households can be categorized as having "poor" (0-28), "borderline" (28.5-42), "acceptable low" (42.5-52) or "acceptable high" (≥ 52) food consumption[6].

Table 1
Categories of the Food Consumption Score

Category	Weight	Detail
Staples	2	rice, cereals and tubers
Pulses	3	lentils and nuts
Vegetables	1	green and leafy vegetables
Fruits	1	fruits
Meat and fish	4	meat, fish and poultry
Dairy	4	milk and butter
Sugar and Sweets	0.5	sugar
Flour	0.5	wheat flour

Controls

Food Consumption and Household Dietary Diversity Scores correlate with family wealth; moreover, the baseline round contains detailed questions on household and agricultural assets. To measure asset ownership inequality, I constructed a measure (index) from this data, as pointed out by McKenzie (2005) and Vyas and Kumaranayake (2006); there are two competing approaches to construct this measure, one consists of adding every asset and then

[6]The last category is usually defined as "oil" but since there was no question in the survey it was replaced for its closest counterpart, wheat flour.

use the total count as the measure, ignoring the explicit heterogeneous nature of every asset (e.g., plows, tractors, carts). The second approach depends on the first component from principal components analysis (PCA) as a proxy for asset ownership; since all asset variables are dummies (equal to 1 if household h owns asset p), then the first principal component is the linear combination:

$$PC_1 = \alpha_1 \left(\frac{x_1 - \overline{x}_1}{s_1} \right) + \alpha_2 \left(\frac{x_2 - \overline{x}_2}{s_2} \right) + \cdots + \alpha_p \left(\frac{x_p - \overline{x}_p}{s_p} \right)$$

And magnitudes α_k / s_k give the effect of a change from 0 to 1 in x_k on the first principal component PC_1, this index provides maximum discrimination between households, given a much larger weight to assets which vary most across households (if all household own asset k, then its weight α_k will be zero, in a similar manner, if no household owns asset k, its weight will also be zero). This measure of asset ownership provides a measure of wealth inequality that is based on a single variable per household that can be used in a regression framework and is a proxy for the standard of living (wealth).

I calculated two indexes from the information gathered in the midline survey: a household asset index for all assets that can be used to proxy for wealth and an agricultural asset index to account for the complexity of their farm operation. To control for baseline characteristics, I use the same set of variables detailed in Shukla and Baylis (2019): age and educational attainment of household head, size of household, presence of migrants, farm size[7], number of plots, a loan dummy equal to 1 if the farmer took any loans, and a measure of per capita food expenditure.

The measure of per capita food expenditure is defined as an equivalence scale (Deaton, 2018, p. 243) to correct for the fact that children have different nutritional requirements than adults; for this purpose I used the "OECD -modified scale"[8] which assigns weights equal to 1 for the first adult, 0.5 for the subsequent adults and 0.3 for children (persons ages 16 or younger).

Self-reported flood damage

Flood experience is self-reported and reflects a case-by-case response; in the midline survey, farmers reported whether floods (here defined as "excessive rain") had any impact on their crops over the last year, out of 3869 total individuals, only 333 answered "Yes" (8.61%). Moreover, 210 of these 333 affirmative answers were from one district (Banka). This pattern is not expected since lands closer to significant rivers are the ones that are often flooded during the rainy season. Still, Banka district is located in the southeast of Bihar. I argue that there is a significant under-report of floods and/or flood damage in the survey, as suggested by Guiteras et al. (2015) since farmers in Bihar are likely to be very adapted to the climate patterns in the region, implying that their answers are influenced by their degree of adaptation.

I can then replace this self-reported impact data with a satellite-measured flood extent, but there exists another issue, in midline and endline rounds, enumerators recorded the

[7] a categorical variable equal to "landless" (if a farmer owns no land), "marginal" (if a farmer owns less than 1 hectare), "small" (if a farmer owns between 1 and 2 hectares) and "large" (if a farmer owns more than 2 hectares)

[8] http://www.oecd.org/els/soc/OECD-Note-EquivalenceScales.pdf

coordinates of each household using hand-held GNSS/GPS devices; then, I can geolocate every household where interviews were carried out with these latitude-longitude pairs, but I am not able to obtain the locations of farms, this is an issue since most farmers live in small or medium-sized towns.

A solution to this problem involves defining a "buffer" (Bolstad, 2016, p. 396) (Bivand et al., 2008, Chapter 4), a circular area with a specified radius around each spatial point, where I will assume the farm must be located, but since this radius is entirely arbitrary, I used three different buffers with radii equal to 500 meters, 1.5 kilometers and 2.5 kilometers[9], so that a farm is classified as flooded if the intersection between any of those buffers and any flooded pixel is non-empty.

There is an implicit trade-off between precision and confidence in buffer radius decision; the more significant the buffer zone, the more likely it is to classify a farm as flooded while increasing the misclassification likelihood. Once these three buffers were calculated around every latitude/longitude pair, a "spatial join" operation was performed with the elevation, rainfall, and distance to water sources data (Kudamatsu, 2018) to construct the final data set.

Identification Strategy

The identification strategy used to identify causal effects strongly depends on the random variation of weather patterns across locations, call F_d our measure of floods, for example, a dummy equal to 1 if a district d was flooded in the year before the survey round was conducted, then, the direct impact of floods on an outcome Y_i can be estimated as:

$$Y_{id} = \alpha + \beta F_{id} + \gamma \mathbf{X}_{id} + \epsilon_{id}$$

Then, the **Average Treatment Effect** (Deryugina and Hsiang, 2014, Hsiang, 2016) of floods will be:

$$ATE = \beta = E[Y_{id}/F_{id} = 1, \mathbf{X}_{id}] - E[Y_{id}/F_{id} = 0, \mathbf{X}_{id}]$$

Since I cannot observe a district with and without floods in the same period, I could use observed values for any other district $j \neq i$ such that an estimator for the ATE can be constructed in the following manner:

$$\hat{\beta} = E[Y_{jd}/F_{jd} = 1, \mathbf{X}_{jd}] - E[Y_{id}/F_{id} = 0, \mathbf{X}_{id}]$$

This estimated ATE will be equal to its actual value if and only if the **Unit Homogeneity Assumption** holds:

$$E[Y_{id}/F_{id}, \mathbf{X}_{id}] = E[Y_{jd}/F_{jd}, \mathbf{X}_{jd}]$$

This is the same as assuming that floods are randomly assigned to geographical units. It can be thought of as a causal analog of early climate impact models such as the "Ricardian approach" pioneered by Mendelsohn et al. (1994), who measured the elasticity of land prices to changes in location, land characteristics, and climate using a hedonic pricing model.

[9] The calculation of these radii was performed using two R packages: *raster* and *sp* (Pebesma, 2018).

My cross-sectional causal approach is vulnerable to omitted variable bias; there is no possible way to ascertain which variables should be included in $\mathbf{X}$. Similarly, it cannot be assumed that their relationship must be necessarily linear either, as pointed out by Lewbel (2019); one alternative is following Mendelsohn et al. (1994) and saturate the model with interactions of control variables; another possibility is to model the interaction using flexible functional forms in a semi-parametric fashion, e. g. such as:

$$Y_{id} = \alpha + \beta F_{id} + g\left(\mathbf{X}_{id}\right) + \epsilon_{id}$$

Where $g(.)$ is a nonlinear function (Li and Racine, 2006, Chapter 7) that can be estimated using splines of Generalized additive models.

Models

Flood reports

The first research question I want to answer is whether there is any evidence of flood under-reporting; given the form of my data, it is better to use the empirical set up of Guiteras et al. (2015) and estimate the probability that an individual i, living in district d reports experiencing a flood:

$$F_{id}^{report} = \Phi\left(\beta_0 + \beta_1 F_{id} + \beta_2 precip_d + \beta_3 alt_{id} + \beta_4 dist_i + \mathbf{X}_{id}\delta\right) \tag{1}$$

$\mathbf{X}_{id}$ is a vector of individual characteristics, F_{id} is a dummy indicating whether household i in district d lives around a radius with flooded areas indicated by satellite pictures, $precip_d$ is the deviation of rainfall from its long-term mean, $dist_i$ is the distance of household i to the nearest water source (rivers or lakes). $\Phi(.)$ is the link function, the cumulative standard normal distribution; this specification allows estimation of the self-reporting bias at the individual level. alt_{id} is a variable measuring the elevation of a specific farm, Kocornik-Mina et al. (2020) uses a dummy that is equal to 1 if the elevation of farm i is less than 10 meters, and I will adopt a similar approach.

The midline round asked farmers if they experienced any unexpected losses due to "heavy rainfall/drought or flood in the past two years (2015 and 2016)". As shown in a previous section, total precipitation in 2015 was around its average, while 2016 total precipitation was above average. Thus most of these losses should come from floods rather than droughts. I depart from Guiteras et al. (2015) in two aspects, first, I include data on geographical features (altitude, distance to rivers), interactions, and a set of controls, and second, I cluster standard errors at the village level to control for spillover effects, if this village-level clustering is not taken into account, standard errors are very small, leading to under-rejection of the null hypothesis of no effect of floods.

Floods and food security

The extent to which floods affect food security outcomes (HDDS and FCS) can be estimated similarly to Le (2020), Nguyen and Nguyen (2020) as:

$$Y_{it} = \alpha_0 + \alpha_1 F_{it} + \gamma \mathbf{X}_i + \gamma_d + \rho_t + \epsilon_{id} \tag{2}$$

Where Y_{idt} are the two food security outcomes (HDDS and FCS), F_{id} is a dummy equal to 1 if household i's buffer intersects with at least one flooded area as detected by remote sensors, γ_d are district-level fixed effects and ρ_t are round fixed effects and $\mathbf{X}_i$ is a vector of baseline characteristics (age of household head, educational attainment of household head, household size, presence of migrants, farm size categories, number of plots, any loans taken), and two variables from the midline round: household asset and agricultural asset indexes.

This cross-sectional analysis allows for estimating the differential impacts of weather shocks on a series of outcomes measured during the survey. I can measure household resilience to these shocks by looking at their effects on either income or consumption; for the former, I can use a measure of total production for every farmer or a measure of monetary income; for the latter, the Food Consumption Score can be used as described by Shukla et al. (2023), to account for observable differences in consumption patterns across households.

As mentioned earlier, the Food Consumption Score takes two forms: a discrete scale taking values from 0 to 112[10], and the second one is the ordinal scale detailed earlier. The first form can be incorporated into the linear model described in the last paragraph; however, the second form requires a different approach, one way to correct this problem is to use an **Ordered Response Model** (Wooldridge, 2010, p. 165).

An ordered choice model such as this can better capture revealed preferences because they are a mapping from an ordered set of preferences into an observed ordinal scale (Greene, 2018, p. 868). Let FCS^* be a variable containing the four FCS categories ("poor", "borderline", "acceptable low", and "acceptable high"), then the estimated model is:

$$FCS^*_{it} = \mathbf{x}'_{it}\beta + \eta_{it} \tag{3}$$

And let $\alpha_1 < \alpha_2 < \alpha_3$ be the cut points (which will be treated as unknown) and define:

$$FCS = \begin{cases} 0 & FCS^* \leq \alpha_1 \\ 1 & \alpha_1 < FCS^* \leq \alpha_2 \\ 2 & \alpha_2 < FCS^* \leq \alpha_3 \\ 3 & FCS^* \geq \alpha_3 \end{cases}$$

Then for $j = \{0, 1, 2, 3\}$ (response categories):

$$\begin{aligned} \Pr(FCS = j) &= \Pr(\alpha_{j-1} < FCS^* \leq \alpha_j) \\ &= \Pr(\alpha_{j-1} - \mathbf{x}'_{it}\beta < \eta_{it} \leq \alpha_j - \mathbf{x}'_{it}\beta) \\ &= \Phi(\alpha_{j-1} - \mathbf{x}'_{it}\beta) - \Phi(\alpha_j - \mathbf{x}'_{it}\beta) \end{aligned}$$

Where $\Phi(.)$ is the cumulative normal standard distribution function. Just like any probit model, this model does not have a meaningful conditional mean function, and as a consequence, the estimation of marginal effects is not straightforward; an expression for the marginal effect of response category j is:

$$\frac{\partial \Pr(FCS = j)}{\partial x_k} = \left[\phi(\alpha_{j-1} - \mathbf{x}'_{it}\beta) - \phi(\alpha_j - \mathbf{x}'_{it}\beta)\right] \beta_k$$

[10] In this sample, the lowest value is equal to 9

$\phi(.)$ is the standard normal density function; finally, given that there are as many marginal effects as observations, I use the average partial effect (APE) as a summary statistic:

$$APE_k = N^{-1} \sum_{i=1}^{N} \left[\phi(\alpha_{j-1} - \mathbf{x}'_{it}\beta) - \phi(\alpha_j - \mathbf{x}'_{it}\beta) \right] \beta_k$$

However, since my variable of interest (being affected by floods) is a dummy, its APE is (Greene, 2018):

$$APE_{flood} = N^{-1} \sum_{i=1}^{N} \Big(\Big[\phi(\alpha_{j-1} - \beta^{flood}) - \phi(\alpha_j - \beta^{flood}) \Big] - \big[\phi(\alpha_{j-1}) - \phi(\alpha_j) \big] \Big)$$

Results

Impact of floods on reports

Table 2
Floods: Self-reports versus satellite data

	(1)	(2)	(3)
Floods (500m)	-0.766 $(0.280)^{**}$		
Floods (1500m)		-0.509 $(0.251)^{*}$	
Floods (2500m)			-0.642 $(0.213)^{**}$
Altitude (SRTM)	-0.017 $(0.002)^{***}$	-0.017 $(0.002)^{***}$	-0.016 $(0.002)^{***}$
Distance to water sources	0.000 $(0.000)^{***}$	0.000 $(0.000)^{**}$	0.000 $(0.000)^{*}$
Total rainfall (Kharif 2015 and 2016)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Controls	Yes	Yes	Yes
McFadden's R^2	0.343	0.342	0.355
AIC	1470.509	1472.598	1443.644
BIC	1569.910	1571.999	1543.045
Log Likelihood	-719.254	-720.299	-705.822
Deviance	1438.509	1440.598	1411.644
Num. obs.	3687	3687	3687

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table 2 shows estimated parameters for equation 1 using all three buffer radii and with the corresponding standard errors clustered at the village level. First, coefficients for flooded areas are all negative across all three specifications, meaning that farmers in flooded areas are less likely to report damages from floods than those in non-flooded areas; this evidence follows the same lines as Guiteras et al. (2015), farmers are less likely to report flood damage due to either a greater adaptation to their environment, e.g., they take for granted that floods will reduce their harvest, farmers may also engage in avoidance behavior (Graff Zivin and Neidell, 2013) taking necessary steps to avert damage from floods, such as storing grains in higher places or selling their grains right after the harvest and buying those same grains after the monsoon season for a higher price to use them as food or animal feed[11].

[11] The literature calls this phenomenon "sell low-buy high".

The negative coefficient on the SRTM-estimated altitude has a similar interpretation since lower elevations positively correlate with flooding chances; farmers living in lower areas are also less likely to report floods. Notice that I used the estimated altitude of their household rather than a buffer since there are no significant variations in altitude across Bihar (see Figure 3). My measure of altitude is obtained via extrapolation.

Distance to water sources positively correlates with the likelihood of reporting floods; farmers closer to water sources are more likely to report damage from floods. However, the significance of this coefficient has to be carefully interpreted since this variable has a high positive correlation with the flood indicator and a negative correlation with altitude.

Finally, rainfall has no significant relation with flood reports; this reflects the fact that both monsoon rainfall and ice melting in the Himalayas are the sources of floods in the Indo-Gangetic plains, both taking place during spring/summer in a predictable pattern, an idea also stressed out by Guiteras et al. (2015) who argue that rainfall is a poor proxy for floods because they (floods) are a consequence of a complex set of hydrological conditions in an area.

It is important to stress that standard error clustering at the village level significantly decreases the significance of all coefficients; had they not been corrected, they would be significant at the 1% level, Guiteras et al. (2015) does not perform such correction (this is clear from the code used to generate the tables), these results suggest lack of robustness to clustering.

Table 3
Average Partial Effects of floods on self-reported damage

	(1)	(2)	(3)
Floods (500m)	$-0.121\ (0.057)^{*}$		
Floods (1500m)		$-0.060\ (0.033)$	
Floods (2500m)			$-0.071\ (0.027)^{**}$
Altitude (SRTM)	$-0.002\ (0.000)^{***}$	$-0.002\ (0.000)^{***}$	$-0.002\ (0.000)^{***}$
Distance to water sources	$0.000\ (0.000)^{**}$	$0.000\ (0.000)^{*}$	$0.000\ (0.000)^{*}$
Total rainfall (Kharif 2015 and 2016)	$0.000\ (0.000)$	$0.000\ (0.000)$	$0.000\ (0.000)$
AIC	1480.690	1487.386	1456.235
BIC	1567.666	1574.362	1543.211
Log Likelihood	-726.345	-729.693	-714.118
Deviance	1452.690	1459.386	1428.235
Num. obs.	3687	3687	3687

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table 3 shows the Average Partial Effects at the median of marginal effects ($\phi(x_i'\hat{\beta})$), defined as:

$$APE = \frac{\partial \Pr\left(F_i^{report}/\mathbf{X}\right)}{\partial F_i} = N^{-1}\sum_{i=1}^{N}\phi(x_i'\hat{\beta})\hat{\beta}_i^F$$

That is, the sample average of marginal effects, the coefficients of distance to water sources, and precipitation are (on average) zero, thus showing an insignificant impact of both variables on the probability of reporting floods.

The altitude coefficient is still tiny but is negative when evaluated at the average marginal effect. Finally, all flood indicator coefficients are negative but either insignificant or marginally

significant, hinting at the existence of heterogeneity across the distribution of marginal effects.

Impact of floods on food security outcomes

Table 4
Ordered probit estimation of flood effects on FCS

	Pr(FCS=j)
Flood dummy	$-0.263\ (0.063)^{***}$
Age of HH head	$0.008\ (0.001)^{***}$
Education attainment of HH head	$0.021\ (0.003)^{***}$
Sex of HH head	$0.205\ (0.086)^{*}$
Size of farm	$0.016\ (0.007)^{*}$
Presence of migrants	$-0.223\ (0.039)^{***}$
Ownership: Marginal	$0.169\ (0.033)^{***}$
Ownership: Small	$0.109\ (0.086)$
Ownership: Large	$0.182\ (0.092)^{*}$
HH assets index	$-0.105\ (0.012)^{***}$
Ag assets index	$-0.032\ (0.014)^{*}$
Per capita Food Exp.	$-0.017\ (0.014)$
Threshold (poor->borderline)	$-0.769\ (0.125)^{***}$
Threshold (borderline->acceptable low)	$0.028\ (0.125)$
Threshold (acceptable low->acceptable high)	$0.328\ (0.125)^{**}$
AIC	15077.341
Log Likelihood	−7518.671
Num. obs.	7268
Iterations	5
McFadden's R^2	0.040

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table 4 shows the estimated coefficients for Equation 3 as an ordered probit model with district-level fixed effects and a round fixed effect. To simplify the interpretation of models, I construct a flood dummy variable (F_{it}) equal to 1 if a farmer is contained in at least one of the buffers.

Proximity to flooded areas hurts the average food consumption score, but the magnitude of this coefficient has no meaning; for that reason, I calculated its corresponding marginal effects in Table 5; most coefficients have the expected sign, except for the two asset indexes, which are negative, but not significant.

Table 5
Ordered probit estimation of flood effects on FCS: marginal effects

	Pr("poor")	Pr("borderline")	Pr("acc. low")	Pr("acc. high")
Flood dummy	$0.033\ (0.007)^{***}$	$0.047\ (0.011)^{***}$	$0.014\ (0.004)^{***}$	$-0.096\ (0.021)^{***}$
Age of HH head	$-0.001\ (0.000)^{***}$	$-0.001\ (0.000)^{***}$	$-0.000\ (0.000)^{****}$	$0.003\ (0.000)^{***}$
Educ. att. of HH head	$-0.003\ (0.000)^{***}$	$-0.004\ (0.000)^{***}$	$-0.001\ (0.000)^{***}$	$0.008\ (0.001)^{***}$
Sex of HH head	$-0.004\ (0.02)^{*}$	$-0.037\ (0.015)^{*}$	$-0.008\ (0.002)^{**}$	$-0.079\ (0.034)^{*}$
Size of farm	$-0.002\ (0.001)^{*}$	$-0.003\ (0.001)^{*}$	$-0.000\ (0.000)^{*}$	$0.006\ (0.002)^{*}$
Presence of migrants	$0.037\ (0.007)^{***}$	$0.041\ (0.007)^{***}$	$0.009\ (0.001)^{***}$	$-0.086\ (0.015)^{***}$
HH assets index	$0.016\ (0.002)^{***}$	$0.019\ (0.002)^{***}$	$0.005\ (0.000)^{***}$	$-0.039\ (0.004)^{***}$
Ag assets index	$0.004\ (0.002)^{*}$	$0.005\ (0.002)^{*}$	$0.001\ (0.000)^{*}$	$-0.012\ (0.005)^{*}$
Per capita Food Exp.	$0.002\ (0.002)$	$0.003\ (0.002)$	$0.000\ (0.000)$	$-0.006\ (0.005)$

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Farmers who live in flood-prone areas (as indicated by the buffer) are 3.3% more likely to belong to the "poor" category of the Food Consumption Score, 4.7% more likely to belong to the "borderline" category, 1.4% more likely to belong to the "acceptable low" category and 9.6% less likely to belong into the "acceptable high" category, these results imply that the effect of floods is negative throughout the categories, but not in a linear fashion[12]. Marginal effects show that the negative effect of floods on FCS categories is driven by the effect on the highest category ("acceptable high") because a negative flood shock moves individuals away from this category into all lower categories, mainly into the "borderline", rather than the lowest one.

Table 6
Household Dietary Diversity Score: OLS

	HHDS
Floods dummy	$-0.115\ (0.510)$
Age of HH head	$-0.005\ (0.003)$
Education attainment of HH head	$0.013\ (0.010)$
Sex of HH head	$0.360\ (0.191)$
Size of farm	$0.043\ (0.017)^{*}$
Presence of migrants	$-0.313\ (0.106)^{**}$
Ownership: Marginal	$-0.465\ (0.133)^{***}$
Ownership: Small	$-1.361\ (0.306)^{***}$
Ownership: Large	$-0.195\ (0.288)$
HH assets index	$-0.091\ (0.030)^{**}$
Ag assets index	$0.068\ (0.029)^{*}$
log per capita food exp.	$-0.130\ (0.087)$
Num. obs.	7190
R^2 (full model)	0.054
R^2 (proj model)	0.029
Adj. R^2 (full model)	0.052
Adj. R^2 (proj model)	0.027
Num. groups: district_ID	5
Num. groups: round	2

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

[12]By construction, the sum of marginal effects is equal to zero.

The Household Dietary Diversity Score is an ordered scale that can also be estimated with OLS as Jodlowski et al. (2016) as shown in Equation 2. Table 6 shows that the flood indicator is negative but insignificant, as well as most variables included in the model. There are at least two possible explanations for this result:

1. As shown by the ordered probit model, the impact of floods is heterogeneous across the error distributions. Thus, the impact on the mean HDDS may be rather small and insignificant, as shown by the model.

2. Climate-related variables have nonlinear effects on observable outcomes, as detailed by Hsiang (2016) and Dell et al. (2014); this lack of significance may result from a misspecified model.

Conclusions

This paper is an extension of Shukla and Baylis (2019) and Shukla et al. (2023); my objectives were, first, to study the extent to which farmers tend to misreport the impact of floods on their households and farms, second, to study the impact of floods on food security outcomes using data from an RCT aimed at disentangling the impact of hermetic bag provision on these outcomes.

Farmers who live in flood-prone areas are likely to under-report the occurrence and impact of floods due to cognitive biases (Guiteras et al., 2015); since floods are ubiquitous to monsoon seasons every year, Bihari farmers are more likely to have a different reference point than that of an external observer. This cyclical nature of floods makes farmers prone to considerable recall biases unless they are subject to an exceptionally high or low rainfall season.

My data shows that there is a negative and significant relationship between floods (measured using satellite data) and reported flood damage; this implies that farmers who live within a certain radius of flooded areas are less likely to report flood damage than those who live outside of it. Flood reports are negatively correlated with altitude so if a household is located in a low area, they are also less likely to report flood damage. Finally, distance to water sources and total rainfall variables have a positive but negligible effect on flood reports (as shown by their average partial effects). The significance of all coefficients depends on village-level standard error clustering; my data shows that if the correlation between neighbors is not taken into account, standard errors are more minor, thus leading to under-rejection bias, Guiteras et al. (2015) does not take this into account, so I could question whether their results are robust to spatial autocorrelation.

The impact of floods on food security outcomes is small, as expected by the adaptation argument I presented in the previous paragraph; given that floods occur almost every year, farmers are likely to engage in avoidance behavior to protect their harvest and assets, increasing their resilience to climate shocks. An ordered probit regression of the Food Consumption Score indicator on a flood indicator and a series of controls shows that floods negatively affect the probability of having a high score, moving down individuals into the lower categories. However, as expected, the effect is relatively small. On a final note, an OLS regression of the second indicator shows no significant impact on the dietary diversity score.

References

Bandyopadhyay, S. and Skoufias, E. (2015). Rainfall variability, occupational choice, and welfare in rural bangladesh. *Review of Economics of the Household*, 13(3):589–634.

Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). Applied spatial data analysis with r. In Gentleman, R., Hornik, K., and Parmigiani, G., editors, *Use R!* Springer, 233 Spring St. New York, NY.

Bolstad, P. (2016). *GIS Fundamentals: A First Text on Geographic Information Systems*. Elder press, White Bear lake, MN, 5 edition.

Deaton, A. (2018). *The Analysis of Household Surveys*. World Bank Publications. The World Bank.

Dell, M., Jones, B. F., and Olken, B. A. (2014). What do we learn from the weather? the new climate-economy literature. *Journal of Economic Literature*, 52(3):740–98.

Deryugina, T. and Hsiang, S. M. (2014). Does the environment still matter? daily temperature and income in the united states. Working Paper 20750, National Bureau of Economic Research.

Donaldson, D. and Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–98.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, 45(2).

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2.

Graff Zivin, J. and Neidell, M. (2013). Environment, health, and human capital. *Journal of Economic Literature*, 51(3):689–730.

Greene, W. H. (2018). *Econometric Analysis*. Pearson Education, New York, NY, 8 edition.

Guiteras, R., Jina, A., and Mobarak, A. M. (2015). Satellites, self-reports, and submersion: Exposure to floods in bangladesh. *American Economic Review*, 105(5):232–36.

Hsiang, S. (2016). Climate econometrics. *Annual Review of Resource Economics*, 8(1):43–75.

Jain, M. (2020). The Benefits and Pitfalls of Using Satellite Data for Causal Inference. *Review of Environmental Economics and Policy*, 14(1):157–169.

Jodlowski, M., Winter-Nelson, A., Baylis, K., and Goldsmith, P. D. (2016). Milk in the data: Food security impacts from a livestock field experiment in zambia. *World Development*, 77:99 – 114.

Jones, A. D., Shrinivas, A., and Bezner-Kerr, R. (2014). Farm production diversity is associated with greater household dietary diversity in malawi: Findings from nationally representative data. *Food Policy*, 46:1 – 12.

Kennedy, G., Ballard, T., and Dop, M.-C. (2010). Guidelines for measuring household and individual dietary diversity. Technical report, Food and Agriculture Organization of the United Nations.

Kocornik-Mina, A., McDermott, T. K. J., Michaels, G., and Rauch, F. (2020). Flooded cities. *American Economic Journal: Applied Economics*, 12(2):35–66.

Kudamatsu, M. (2018). GIS for Credible Identification Strategies in Economics Research. *CESifo Economic Studies*, 64(2):327–338.

Le, T. N. T. (2020). Floods and household welfare: Evidence from southeast asia. *Economics of Disasters and Climate Change*, 4:145 – 170.

Leroy, J. L., Ruel, M., Frongillo, E. A., Harris, J., and Ballard, T. J. (2015). Measuring the food access dimension of food security: A critical review and mapping of indicators. *Food and Nutrition Bulletin*, 36(2):167–195. PMID: 26121701.

Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903.

Li, Q. and Racine, J. S. (2006). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Marchetta, F., Sahn, D. E., and Tiberti, L. (2019). The Role of Weather on Schooling and Work of Young Adults in Madagascar. *American Journal of Agricultural Economics*, 101(4):1203–1227.

McKenzie, D. J. (2005). Measuring inequality with asset indicators. *Journal of Population Economics*, 18:229–260.

Mendelsohn, R., Nordhaus, W. D., and Shaw, D. (1994). The impact of global warming on agriculture: A ricardian analysis. *The American Economic Review*, 84(4):753–771.

Nguyen, G. and Nguyen, T. T. (2020). Exposure to weather shocks: A comparison between self-reported record and extreme weather data. *Economic Analysis and Policy*, 65:117 – 138.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446.

Shukla, P. and Baylis, K. (2019). *Subsidizing Technology Adoption*. PhD thesis, Department of Agricultural and Consumer Economics - University of Illinois at Urbana-Champaign, 1301 W. Gregory Dr. Urbana, IL, United States of America.

Shukla, P., Pullabhotla, H. K., and Baylis, K. (2023). The economics of reducing food losses: Experimental evidence from improved storage technology in india. *Food Policy*, 117:102442.

Spence, A., Poortinga, W., Butler, C., and Pidgeon, N. F. (2011). Perceptions of climate change and willingness to save energy related to flood experience. *Nature Climate Change*, 1(1):46–49.

Swindale, A. and Bilinsky, P. (2006). Household dietary diversity score (hdds) for measurement of household food access: Indicator guide (version 2). Technical report, Food and Nutrition Technical Assistance Project (FANTA), 1825 Connecticut Avenue, NW Washington, D.C. 20009-5721.

Vyas, S. and Kumaranayake, L. (2006). Constructing socio-economic status indices: how to use principal components analysis. *Health Policy and Planning*, 21(6):459–468.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.