Nowcasting disruptions to human capital formation: Evidence from high-frequency household and geospatial data in rural Malawi*

Elizabeth J. Tennant† Aleksandr Michuda‡ Joanna B. Upton†
Andres Chamorro§ Ryan Engstrom¶ Michael L. Mann¶
David Newhouse§ Michael Weber‖ Christopher B. Barrett**

Revised in response to peer review: June 7, 2026

Abstract

Exposure to extreme weather events and other adverse shocks has led to an increasing number of humanitarian crises in developing countries. These events cause acute suffering and compromise future welfare by adversely impacting human capital formation among vulnerable populations. Early and accurate detection of adverse shocks to food security, health, and schooling is critical for timely, targeted humanitarian interventions. Yet monitoring data are rarely available with the frequency and spatial granularity needed. Here, we use high frequency household survey data from the Rapid Feedback Monitoring System (2020-2023) in southern Malawi, to explore whether combining monthly data with publicly available remote-sensing features improves the accuracy of machine learning extrapolations across time and space, thereby enhancing monitoring efforts. In our sample, illnesses and schooling disruptions are not reliably predicted. When both lagged outcome data and geospatial features are available, intertemporal and spatio-temporal prediction of food insecurity indicators is promising.

*Funding: We gratefully acknowledge funding from the World Bank's Research Support Board, Human Capital Project, and the UK Government through the Data and Evidence for Tackling Extreme Poverty (DEEP) Research Program. Acknowledgements: We appreciate the feedback of seminar and conference audiences at the Collaboration for International Development Economics Research (CIDER) and Catholic Relief Services (CRS) Workshop at Cornell, The Pulse of Progress: Harnessing High-Frequency Survey Data for Development Research in the Polycrisis Era Conference in Washington, D.C., and the World Data Forum in Medellin, Colombia. We thank the Malawi RFMS team, including James Campbell, Trovehey Kazunguza, Edward Mtemwende, Chisomo Ngosi, and Patrick Tembo. We also thank Gabriel Demombynes, Olivier Dupriez, Haishan Fu, Keith Garrett, and Craig Hammer for support. We are grateful to the Editor and the Reviewers for their careful evaluation of our work and for their thoughtful feedback.

†Charles H. Dyson School, Cornell University, Ithaca, NY 14853
‡Department of Economics, Swarthmore College, Swarthmore, PA 19081
§Development Economics Data Group, World Bank Group, Washington DC 20433
¶Department of Geography, The George Washington University, Washington DC 20052
‖Human Capital Project, World Bank Group, Washington DC 20433
**Charles H. Dyson School of Applied Economics and Management and Jeb E. Brooks School of Public Policy, Cornell University, Ithaca, NY 14853

1 Introduction

Rural households in low- and middle-income countries (LMICs) are exposed to a wide variety of adverse shocks, including weather extremes, disease and pest outbreaks, price fluctuations, and job losses. These events cause acute suffering and often compromise future welfare by disrupting human capital formation among vulnerable populations who lack the resources or social safety nets to cushion these shocks. There is strong evidence that even short- to medium-duration disruptions to nutrition, health, and education—such as those induced by droughts, winds, or flooding—can have lasting and even inter-generational consequences for human capital, thereby adversely affecting long-run prospects for inclusive economic growth (Hoddinott and Kinsey, 2001; Victora et al., 2008; Adair et al., 2013; Acemoglu et al., 2014; Fishman et al., 2019; Rossi, 2020; Anttila-Hughes et al., 2021; Blom et al., 2022; Rossi and Weber, 2024).

Improved real-time monitoring of food security, health, and education indicators can assist in detecting when and where human capital formation might be disrupted, helping to guide interventions that prevent unnecessary suffering and long-term harm. When combined with information on current human capital stocks and the infrastructure that supports their development—such as food systems, health services, and education—this can help to guide interventions that prevent unnecessary suffering and long-term harm. However, high-quality, up-to-date household survey data with broad geographic coverage are rare. Established household surveys, such as the Demographic and Health Surveys (DHS) and the Living Standards Measurement Study (LSMS), provide a rich picture of health and well-being, but are collected infrequently and published with significant lags. These data may therefore overlook critical periods or be published too late for use in responding to crises as they unfold. High-frequency household surveys are emerging as one way to more nimbly detect trends in socioeconomic indicators, but such data remain scarce and methods for using them are still developing (Gourlay et al., 2021; Abay et al., 2022; Abate et al., 2023; Dillon et al., 2025).

Where costs, capacity, or access constrain the availability of survey-based data, one way forward is data fusion: integrating complementary data streams to construct a more complete picture. Global institutions routinely employ temporal forecasting to fill data gaps, for instance, using aggregate macroeconomic indicators and statistical models to interpolate or extrapolate national labor trends or poverty rates when annual surveys are lacking (Mahler et al., 2022; ILO, 2026).

A growing literature seeks to fill gaps at a much higher spatial resolution by training models to predict well-being indicators, such as poverty, asset wealth, food security, or malnutrition, from remotely sensed or other alternative data sources (Burke et al., 2021; McBride et al., 2021; Aiken et al., 2022; Newhouse, 2024). Most micro modeling efforts concentrate on spatial rather than temporal out-of-sample prediction. The few exceptions that attempt inter-temporal prediction find that predicting change is more difficult than predicting spatial variation, and predominantly train on low-frequency household surveys examining changes over periods of a year or more (Lentz et al., 2019; Browne et al., 2021; Tang et al., 2022; Yeh et al., 2020; Marty and Duhaut, 2024; Zheng et al., 2025). However, this is changing as high-frequency survey data become more available (Mude et al., 2009; Constenla-Villoslada et al., 2025).

Existing cross-sectional modeling shows that the capability of data fusion models varies with the type of outcomes predicted. Asset wealth indices are typically more predictable than measures of malnutrition (Browne et al., 2021) or the flows of consumption expenditures (Yeh et al., 2020; Tennant et al., 2025). Head et al. (2017) used neural networks to predict asset wealth with an out-of-sample R^2 of about 0.70, but the same approach only predicts education levels with $R^2 = 0.47$ and poorly predicts ($R^2 = 0.06$) the child weight-to-height percentile. Performance may also depend on the intended use and evaluation metric. For example, remote-sensing-based poverty mapping in Malawi has been shown to capture broad spatial patterns of poverty but to provide less reliable estimates for individual small areas (van der Weide et al., 2024).

The predictability of asset wealth follows from the relative stability of durable assets and our ability to detect or proxy these features using remote sensing. For example, durable assets such as buildings are directly detectable in satellite imagery, while features like nighttime lights proxy wealth by capturing the extent of electrification and urbanization. Stability is advantageous when our goal is to understand structural differences (Tennant et al., 2025). But it can be a disadvantage for monitoring and early warning, especially during moments of disequilibrium brought on by shocks, when the goal is to intervene before people disinvest in human capital and resort to distress sales of their productive assets. Efforts to date have achieved some success at aggregate scales (Andree et al., 2020) or when models incorporate lagged survey data (Lentz et al., 2019; Constenla-Villoslada et al., 2025). Predicting at high spatial resolution for un-surveyed areas remains an unresolved challenge. Here, the modeler must capture aspects of well-being that are inherently less stable, more sensitive to seasonal stressors and stochastic shocks, and that arise from underlying processes that may be more difficult to proxy with satellite imagery. These difficult-to-predict outcomes are among the most critical to early warning systems.

1.1 Early Warning and Human Capital in Malawi

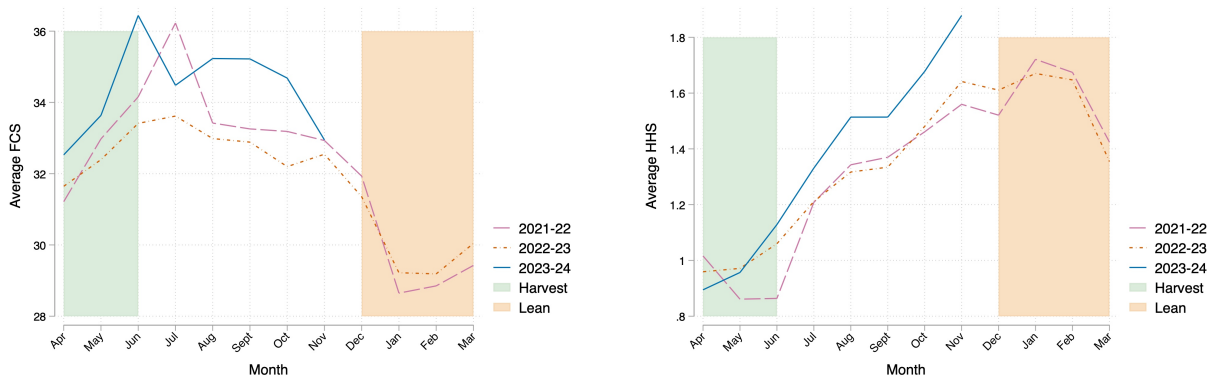
In Malawi and many LMICs, the mobilization of national safety nets and international humanitarian assistance for food crises is heavily dependent on the Integrated Phase Classification (IPC) system. While coordinated globally, this framework is operationalized locally—in Malawi, under the leadership of the national Vulnerability Assessment Committee (MVAC) in close coordination with the Famine Early Warning Systems Network (FEWS NET) and other national and international partners. To assess and forecast the severity of acute and chronic food insecurity, multi-stakeholder working groups rely on a convergence-of-evidence approach to triangulate available data and models, including field surveys, market assessments, and seasonal crop estimates. This system has led to critical improvements in the detection of broad regional crises and the coordination of humanitarian responses. However,

IPC assessments and their underlying data systems operate primarily at the livelihood-zone or district level, are published with limited frequency, and may systematically under-estimate true levels of food insecurity—possibly due to political elements of the consensus process (Maxwell and Hailey, 2021; Lentz and Maxwell, 2022; Barrett, 2025; Lentz et al., 2025). Furthermore, because these systems focus predominantly on food security—with illnesses and schooling falling under the purview of separate, though coordinated, response systems—decision-makers often lack current, comprehensive, and spatially granular estimates of shocks to human capital formation.

The need for timely and localized monitoring is underscored by the fragility of human capital formation in the region. Children born in Malawi in 2020 are expected to achieve only 41% of their full productive potential by adulthood (World Bank, 2021). While Malawi's Human Capital Index (HCI) has improved over the past decade through gains in infant health services, 39% of children under five are stunted, 5% do not survive to age five, and the average child is expected to complete just 5.5 learning-adjusted years of schooling (World Bank, 2021).

Progress in health and learning suffers frequent disruption from predictable seasonal stressors and idiosyncratic shocks. Food security is closely linked to the seasonality of rainfed agriculture, as 85% of Malawians engage in smallholder cultivation of maize (NSO, 2020). Each year during the rainy season, food insecurity climbs as maize stocks are exhausted and prices peak before the main harvest (Ellis and Manda, 2012; Gilbert et al., 2017; Anderson et al., 2018). Figure 1 depicts how, on average, households in our sample exhibit peak food insecurity, with low dietary diversity (panel a) and an increase in adverse coping strategies (panel b) in January and February before the start of the green harvest in March (see Online Appendix Figures S2 and S3 for trends in illness and schooling disruptions).

The impact of these cyclical stressors varies depending on vulnerability and unpredictable shocks, such as extreme weather. As a result, Malawi exhibits not only seasonal fluctuations but also significant year-to-year, seasonally-adjusted variation in food insecurity and disease



(a) dietary diversity

(b) adverse coping strategies

Figure 1: Average food (in)security over time in the survey sample. Panel (a) plots the Food Consumption Score, an indicator of dietary diversity for which higher scores are 'better' (more food secure). Panel (b) plots the Household Hunger Scale, a measure of households' use of adverse coping strategies for which lower is 'better' (more food secure).

burden. Maize is drought-sensitive, yet few Malawian smallholders have access to irrigation. Dry spells can dramatically reduce yields, while heavy rainfall late in the growing season can damage promising harvests.

Disease patterns and schooling disruptions may be particularly challenging to predict because they reflect complex interactions of climate, infrastructure, and human behavior, and may be pathogen specific. Tuberculosis, malaria, and cholera exhibit distinct seasonal or weather-driven patterns (Kirolos et al., 2021; Hajison et al., 2017; Miggo et al., 2023). These complex drivers may help explain why self-reports of illness show no single clear seasonal pattern (Online Appendix Figure S2) and why overall disease prevalence proves difficult to model accurately. A further complication is that our training period overlaps with COVID-19 border and school closures (2020–2021), which likely introduced atypical variance into our illness and schooling disruption metrics (Upton et al., 2023).

Adverse weather shocks in Malawi and neighboring countries have been shown to reduce school enrollment, though effects may be partially offset by safety nets such as school feeding programs (Staffieri et al., 2023; Beegle et al., 2006; Agamile and Lawson, 2021; Glick et al., 2016). These findings are based on low-frequency surveys and cannot identify seasonal com-

ponents. The survey data (described below) show more inter-annual than seasonal variation in schooling interruptions due to shocks (Online Appendix Figure S3).

How households choose to allocate resources during times of scarcity shapes the downstream effects of shocks on human capital investment, which may explain the heterogeneous findings regarding how shocks impact the health and education of different sub-groups (Maccini and Yang, 2009; Anttila-Hughes and Hsiang, 2013; Björkman-Nyqvist, 2013; Glick et al., 2016; Staffieri et al., 2023). In a data fusion context, these highly context-dependent coping strategies introduce significant statistical noise; more heterogeneous relationships may be inherently difficult for an algorithm to learn from geospatial features—particularly when training samples are limited.

1.2 Nowcasting disruptions to human capital formation

The motivation for this paper is to improve real-time monitoring and early warning systems in Malawi and in other rural, low-income settings. We investigate whether high frequency surveys and publicly available geospatial data improve predictions of when and where communities experience adverse impacts on human capital accumulation at the village level: periods of food insecurity, illness, and schooling disruptions. Our approach combines high-frequency longitudinal household survey data from southern Malawi, interpretable and contextual Earth Observation (EO) based features, and machine learning models to nowcast these outcomes. We simulate model performance under different configurations of survey data availability—each analogous to a real-world decision context—by varying the spatial overlap between training and testing samples and the temporal forecasting horizons.

The training and testing data are drawn from a novel 28 (to 39) month panel dataset of 4,500 households in ten districts of rural southern Malawi, collected from 2020 to 2023 as part of the Rapid Feedback Monitoring System (RFMS) project. The data are collected by trained local enumerators and include monthly measures of food insecurity (including dietary diversity and negative coping strategies), illness, and disruptions to children's schooling.

We match the RFMS data to two sets of geospatial features. The first is a set of interpretable features that correlate with or condition shocks to human capital formation, such as weather, vegetation indices, nighttime luminosity, and measures of building size and density. For example, we hypothesize that anomalous precipitation, heat stress, and cyclone winds affect outcomes through variation in agricultural yields, while building density and nighttime lights proxy economic activity. The second set of geospatial features are more abstract contextual features that statistically quantify edge patterns, textures, and spectral signatures (Graesser et al., 2012; Chao et al., 2021; Engstrom et al., 2015). In a rural setting, these have not previously been tested: we hypothesize that they may help identify physical shocks that affect household outcomes, including crop failure or flood damage.

Using model and cross-validation variants, we explore a range of policy-informed 'information scenarios' that simulate variations in the spatial and temporal availability of survey-based data. First, we predict *temporally* out-of-sample, testing whether we can forecast into periods for which we have only remotely-sensed data and lagged, but not current, survey data. This approach is useful when up-to-date estimates are needed but the most recent survey data are outdated, for example, in the aftermath of an unexpected disruption. Second, we predict *spatially* out-of-sample, simulating a scenario in which we have only a single round of survey data and wish to extrapolate results to non-sampled enumeration areas (EAs). This is analogous to a setting where high-frequency survey data are entirely absent, and outcomes are predicted purely from geospatial data. Our final set of models test whether we can predict both *spatially* and *temporally* outside the training sample, nowcasting into non-sampled districts for which remotely-sensed data are available but household survey data are either unavailable or limited. This simulates the task of predicting outcomes in villages with only a single round of older data, such as from a previous monitoring or household budget survey, using a model trained on high-frequency data from adjacent districts. For all three questions, we explore the sensitivity of predictive accuracy to the duration and recency of the training data, and to the season of prediction.

These models shed light on the data requirements for effective nowcasting, including the extent to which data fusion efforts can leverage an emerging resource: short, monthly household surveys that monitor food insecurity and other indicators of human capital formation. A key contribution of this work is to assess the predictive value of remote sensing indicators for short-term, dynamic outcomes—an application distinct from their more common use in predicting structural or long-term aspects of well-being. We show how these geospatial features, when combined with high-frequency survey data, can support village-level prediction of human capital dynamics in poor, rural settings.

2 Data

2.1 The Rapid Feedback Monitoring System (RFMS)

Outcome indicators are constructed from the RFMS monthly household panel survey, collected in-person by local enumerators from 2020-2023. We utilize data that are representative of 10 districts that cover most of Malawi's southern region, selected based on their classification as the most food insecure. Data collection started first with six districts in August 2020, with four additional districts added in July 2021. Our final sample includes 178 enumeration areas (EAs) of 25 households each, for a total of sample of approximately 4,450 households. The sampling frame was designed in collaboration with the Government of Malawi National Statistics office, to be consistent with national panel surveys. The details of the RFMS sampling frame, regional coverage, attrition, and our sample selection, as well as graphic descriptions of our outcome indicators, are provided in Section 6.1 of the Online Appendix.

To align with the goal of monitoring community-level crises rather than targeting individual households (McBride et al., 2021), we aggregate all household-level survey responses to the EA level (approximately the traditional village group) prior to model training. This aggregation focuses the predictive models on covariate spatial and temporal shocks.

The monthly surveys track food security indicators, both dietary diversity and coping

strategies, and reports of shocks experienced. Shock reports include illnesses within the household, with follow up questions as to the nature of the illness and which household member(s) was(were) sick. We also ask what responses or coping strategies were used in response to shocks, including withdrawing children from school. In Table 1, we describe how the five outcome indicators in this analysis are derived from these monthly questions, with further detail on the methodology provided in Appendix Table S1.

Table 1: Dependent variables constructed from the RFMS data

Variable	Description
Food Consumption Score (FCS)	Food groups consumed in past 7 days, weighted by quality and frequency (Wiesmann et al., 2009)
Household Hunger Scale (HHS)	Weighted sum of three extreme coping strategies, e.g. going a full day without food (Ballard et al., 2011)
Illness, household	Presence of any illness in household
Illness, children 0-5	Presence of child 0-5 with illness in household
Schooling disruption	Household reported removing child from school (short or long term) due to shock of any kind

Notes: For analysis, all indicators are aggregated by taking the EA level mean.

2.2 Geospatial features

The geospatial features are derived entirely from publicly available data and methods. These include interpretable features that capture or proxy phenomena that we expect to correlate with or condition shocks to human capital. We also include more abstract contextual features that statistically quantify the "edge patterns, pixel groups, gaps, textures, and the raw spectral signatures [...] over groups of pixels or neighborhoods" (Chao et al., 2021). All features were extracted at the EA level; the construction and data sources are detailed in the Online Appendix (Tables S3 and S4).

A key challenge when analyzing monthly or quarterly geospatial data is cloud cover.

Malawi experiences persistent cloud cover during the rainy season (especially December to March). The secondary products we utilize account for this in different ways; when direct image quality data is available we include these measures of cloud cover directly in our models. For the contextual features, many of which are sensitive to missing values, we linearly interpolate missing pixels and also include a cloud cover variable.

2.2.1 Interpretable features

We include monthly, interpretable features selected for their potential to either causally influence or proxy the levels and changes to agriculture and livestock conditions, disease environments, and other sources of change or disruption. Our interpretable features are designed to reflect key mechanisms linking geospatial characteristics to outcomes of interest, with a primary focus on inputs and proxies for agricultural productivity. In rural Malawi, households depend heavily on rainfed agriculture—particularly the drought-sensitive staple maize—for both subsistence and income. This reliance makes them highly vulnerable to weather shocks such as dry spells, heat stress, floods, and cyclone winds, which may also be proxied by vegetation indices. Including indicators of differential vulnerability, such as local geography, hydrology, and climatology, can help the model learn and predict the impacts of these shocks on production more effectively.

A detailed description of the variables and the data sources from which they were obtained are described in Panel A of Online Appendix Table S3. We include monthly absolute and relative (local, seasonally adjusted z-scores) measures of rainfall and heat stress, as well as time-invariant indicators of the historical climatology of rainfall and temperature at each location. The Normalized Difference Vegetation Index (NDVI), which has previously demonstrated predictive capacity for inter-temporal changes in consumption in low-income countries including Malawi (Tang et al., 2022), is computed from three different imagery sources: the Moderate Resolution Imaging Spectroradiometer (MODIS), publicly available mosaics by Planet Labs (Pandey et al., 2023), and from Sentinel-2 (see contextual features).

We also include nighttime luminosity, among the first adopted satellite-based proxies of levels of wealth or development (Noor et al., 2008; Elvidge et al., 2009), capable of capturing changes in buildings and infrastructure, population, and the severe weather impacts.

We also include slow-moving or time-invariant features likely to be predictive of well-being in cross-section and to condition the relationship between monthly features and human capital flows and shocks. Detailed in Panel B of Online Appendix Table S3, these include features of the topography and hydrology, including a flow accumulation measure to proxy flood risk. We also include measures of the built environment, including the number of buildings and their average size.

2.2.2 Contextual features

Contextual features capture spatial attributes derived from the arrangement and interaction of pixels within an image, allowing them to represent the physical layout and potential functions of those features. While in urban landscapes these features have proven effective at capturing morphological variations and distinguishing socio-economically deprived regions (Graesser et al., 2012; Duque et al., 2015; Engstrom et al., 2015; Chao et al., 2021; D. Matarira and Naidu, 2022), their utility in rural, agricultural contexts using moderate-resolution imagery remains largely untested. We include these features, which we derive from the European Space Agency Sentinel-2 (Level 2A) imagery, to provide initial evidence on their value-add for nowcasting human capital outcomes in rural areas (see Online Appendix Section 6.3 for details on imagery processing and variable construction).

A drawback of contextual features is that they are not interpretable and policy-makers may therefore be hesitant to trust their predictions. However, previous work demonstrates the value of combining interpretable and imagery-based features, including for inter-temporal tasks, to increase the reliability of predictions (Marty and Duhaut, 2024; Zhou et al., 2022). Models can be run both with and without contextual features, allowing practitioners to compare predictions and identify divergences.

3 Methods

How well can we predict adverse shocks to human capital accumulation using machine-learning models trained on a combination of high-frequency (monthly) household panel surveys and publicly available geospatial data? To address this question, we train OLS, Post-LASSO, and XGBoost models to nowcast food insecurity, illness, and schooling disruptions using three different cross-validation strategies designed to assess capacity for spatial and temporal extrapolation from the survey sample. We test predictions for surveyed areas and areas simulated to be un-surveyed (by excluding them from the training data), varying the duration of the training data as well as the lag structure of the models. Our cross-validation strategy aims to replicate three different information scenarios: testing spatial, temporal, and spatio-temporal out of sample performance.

In the following section, we describe how our models and cross-validation strategies relate to the objectives and data constraints that a practitioner might face.

3.1 Models and cross-validation

We begin with a simple cross-sectional model. This allows us to test the extent to which contemporaneous geospatial features Z_{it} (see Section 2.2) predict variation across enumeration areas in each of our human capital inflow measures of food insecurity, illness, and school disruption y_{it} (see 2.1). Using the algorithms described in Section 3.2, for cluster i in a given time period t we estimate a model:

$$y_{it} = f(Z_{it}) + \varepsilon_{it}. \tag{1}$$

We estimate model (1) separately for each monthly period to simulate a scenario where high-frequency data is unavailable. For each period, we evaluate performance using 'leave-one-district-out" (LODO) (or "leave-one-out", LOO, in which each enumeration area is left out in-turn) spatial cross-validation. This tests our basic ability to capture relationships

between our survey-based measures of adverse shocks to human capital and our geospatial feature set in a cross-sectional context.

Next we model the dynamics of our human capital measures, taking advantage of our rich high-frequency panel data to train the models. Model (2) serves as our baseline model. For each outcome indicator y_{it}, we estimate a model of the basic form:

$$y_{it} = f(y_{i,t-s}, Z_{i,t...t-k}, D_m) + \varepsilon_{it}, \tag{2}$$

where $y_{i,t-s}$ is a single s-month lag of the outcome variable; $Z_{i,t...t-k}$ is a vector of earth observation features for the prediction period and each of the prior k months; and a calendar monthly dummy D_m accounts for seasonality. In the baseline analysis, we set $s = 6$ and $k = 6$. This means that we are predicting outcome y six months forward from the last period for which we observe y in the data; put another way, we simulate 'nowcasting' six months ahead of the latest survey data. We also consider $s = 3$ and $s = 12$ month lags to test whether model performance degrades over time. We do not include a cluster-level fixed effect α_i, which would limit spatial extrapolation and, in combination with the lagged outcome, introduce bias due a mechanical correlation between $y_{i,t-s}$ and the residual (Nickell, 1981).

Model (2) is first evaluated using a temporal cross-validation strategy, which simulates prediction for surveyed areas six months beyond any available survey data. Temporal nowcasts may be a useful monitoring tool when high-frequency survey data collection is interrupted, discontinued, or if survey data processing is delayed. In settings with ongoing survey-based monitoring in place (such as the RFMS), survey-based estimates are preferred.

Next, we implement spatio-temporal cross-validation to evaluate our ability to nowcast into EAs or districts that are not currently covered by high-frequency household surveys, using the LODO (or LOO) approach described above. These models assume that a (single) lagged measure of y is available from s months prior to the prediction period: a comparable real-world scenario would arise if we wanted to nowcast into a location where high-frequency

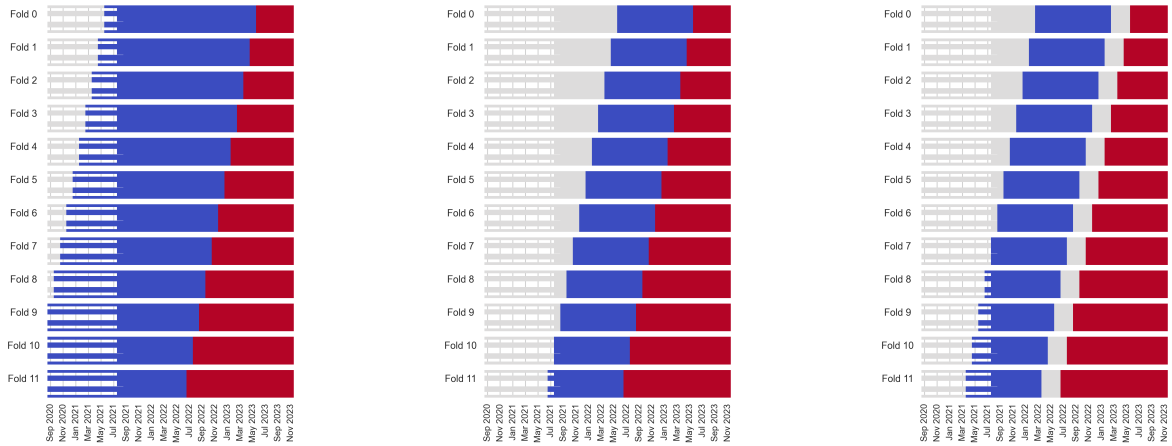
data becomes unavailable, but we do have a recent snapshot of outcome data from a cross-sectional survey. In Malawi, this might include prediction for districts outside the RFMS sample but represented in lower frequency surveys, such as the LSMS, DHS, Multiple Indicator Cluster Survey, or the cross-sectional surveys implemented by the MVAC.

Finally, we train a model that does not rely on prior outcome data, and another that relies only on lagged outcomes. The first simulates prediction into areas for which $y_{i,t-s}$ is not observed, and therefore cannot be on the right-hand side, by estimating $y_{it} = f(Z_{i,t...t-k}, D_m) + \varepsilon_{it}$ (3). We then test how well the high-frequency data performs alone, excluding the geospatial predictors ($Z_{i,t...t-k}$) and estimating $y_{it} = f(y_{i,t-s}, D_m) + \varepsilon_{it}$ (4). Importantly, because model (4) relies solely on lagged survey data, the results are indicative of forecasting skill in contexts where recent lagged outcome data is available. Comparing the results of model (2) and model (4) indicates how much the full set of geospatial indicators contribute to the accuracy of nowcasts.

Our main results depict average outcomes based on different data splits that vary the duration and timing of the training and testing data. This ensures the robustness of our findings, and also allows us to observe if predictive accuracy changes with the duration of the training data or across specific periods of prediction. Figure 2 illustrates the three main typologies that we test for the temporal cross-validation case. In panel (a) we fix the training period at up to 24 months with the test set increasing from 6 to 17 months. Panel (b) is similar to (a), but with a smaller training size of one year. In panel (c) we maintain one year of training data and add an additional gap of three months between training and prediction. In all three cases, we vary the exact timing of the training and prediction periods across eleven folds.

3.2 Algorithms

Before the features are input into each model, they are preprocessed and normalized. Features are imputed using K-nearest neighbor imputation and then scaled by their mean and



(a) Up to Two Years Training, Increasing Test, No Gap

(b) One Year Training, Increasing Test, No Gap

(c) One Year Training, Increasing Test, 3 Month Gap

Figure 2: Temporal Cross-validation Strategies. These panels show the partitioning of training (blue) and testing (red) data under three different temporal cross-validation strategies. In Panel (a), up to two years of temporally-adjacent training data are used; in Panel (b), one year. Panel (c) introduces a temporal gap between training and testing. Note that, as illustrated, the full 1–2 years of training data are not always available for all enumeration areas due to the staggered rollout of the survey.

Table 2: Summary of models and cross-validation variants

Model	**Algorithm(s)**	**Data split(s)**	**Cross-validation**
Cross-sectional (1): $y_{it} = f(Z_{it}) + \varepsilon_{it}$	OLS Post-LASSO XGBoost	District (Alt: EA) *training period:* $T = 1$	Spatial
Dynamic, geospatial & lag (2): $y_{it} = f(y_{i,t-s}, Z_{i,t\ldots t-k}, D_m) + \varepsilon_{it}$, where: $s = 6$ $(3, 12)$ and $k = 6$	OLS Post-LASSO XGBoost	District (Alt: EA) *training period:* $T = 24, 12, 12$+gap	Temporal Spatio-temporal
Dynamic, geospatial only (3): $y_{it} = f(Z_{i,t\ldots t-k}, D_m) + \varepsilon_{it}$ where: $s = 6$ and $k = 6$	OLS Post-LASSO XGBoost	District (Alt: EA) *training period:* $T = 24, 12, 12$+gap	Temporal Spatio-temporal
Dynamic, lag only (4): $y_{it} = f(y_{i,t-s}, D_m) + \varepsilon_{it}$ where: $s = 6$ and $k = 6$	OLS Post-LASSO XGBoost	District (Alt: EA) *training period:* $T = 24, 12, 12$+gap	Temporal Spatio-temporal

Notes: In the above models, described further in the text, y_{it} is the outcome of interest for enumeration area i in time t; $y_{i,t-s}$ is a single s-month lag of the same outcome variable; $Z_{i,t\ldots t-k}$ is a vector of earth observation features for the prediction period and each of the prior k months; and a calendar monthly dummy D_m accounts for seasonality.

variance. Any features with zero variance or at least 80% correlated are dropped.

We estimate the models using ordinary least squares (OLS), post least absolute shrinkage and selection operator (Post-LASSO), and gradient boosting regression, specifically extreme gradient boosting (XGboost). Each model represents a point on the spectrum between interpretability and the ability to capture non-linearity.[1]

The OLS technique for estimating a standard linear regression model is ubiquitous and acts as a suitable baseline method for prediction. OLS is linear in parameters, unable to easily capture complex non-linearity. Additionally, OLS results depend on the choice of model predictors and the method is not suited for ill-conditioned problems, where the number of features exceeds the number of observations.

We also implement the LASSO (and ultimately use it as a part of a Post-LASSO estimation) which is linear, but is robust to ill-conditioning, due to its added regularization constraint. The LASSO algorithm regularizes coefficients in the model using the L1-norm, thus reducing the coefficient on features with limited explanatory power to exactly zero. The LASSO algorithm introduces an added parameter λ which controls how stringently to regularize the coefficients and is tuned from the data. To tune the regularization parameter in the LASSO algorithm we follow Zhang et al. (2010) and choose the λ that minimizes the Bayesian Information Criterion.

After estimating the LASSO algorithm, we re-estimate an OLS model using only the non-regularized features, which are then used for prediction. This regularization and re-estimation procedure is known as Post-LASSO estimation (Belloni and Chernozhukov, 2013). In both of these cases, parameter estimates provide a cursory understanding of how each feature contributes to prediction.[2] Our main results are based on the Post-LASSO specification; however, as a robustness check, we also estimate the LASSO model. The results are

[1] Algorithms were implemented in the Python programming language. OLS was implemented in the `scikit-learn` library. Post-LASSO was implemented in Stata. XGBoost was implemented using the `xgboost` library.

[2] We do not use coefficient magnitudes as a final measure of feature importance, opting instead for calculating Shapley values for each feature and decomposing the values into R^2 contributions (Tallón-Ballesteros and Chen, 2020).

highly similar (Online Appendix Section 8.13).

Finally, we use extreme gradient boosting (XGBoost) as an implementation of gradient boosting trees (Chen and Guestrin, 2016). XGBoost (and gradient boosting more generally) combines the ability of decision tree algorithms to capture non-linearity, and the power of model ensembling for improved and robust prediction. Decision or regression trees learn the data by finding the optimal data "splits" that lead to a correct prediction. Boosting creates high predictive accuracy by starting with a decision tree and iteratively improving accuracy by learning the pseudo-residuals of the tree in the previous iteration. In a sense, boosting helps a machine learning algorithm to "learn" from its mistakes (Friedman, 2002). While gradient tree boosting is a powerful algorithm for machine learning, given that it is a set of decision trees, it has many parameters that need to be tuned to avoid overfitting to the data. See Sections 3.1 and Online Appendix Section 7 for overviews of our approach to training and tuning.

4 Results

Model performance varies substantially depending on the algorithm used, the outcome type, and data availability. The XGBoost and Post-LASSO models show considerable promise for food insecurity monitoring applications. In contrast, predictive performance is consistently weaker for outcomes related to illness and schooling disruptions. Accordingly, we present the results for these categories separately.

On average, Post-LASSO appears to have a slight edge over XGBoost. Smaller sample size may lead to overfitting in the XGBoost models, even after tuning. Although high frequency data provide larger sample sizes than are usually available from household surveys, it may still provide too little data to reliably train a complex tree model such as XGBoost. Since OLS does not include any regularization, the large feature set relative to the smaller sample size results in poorly specified models and inaccurate predictions.

4.1 Nowcasting Food Security Outcomes

In this section, we report model performance in predicting two indicators of food security: the FCS, which captures dietary diversity, and the HHS, which reflects the use of adverse coping strategies. The results indicate that algorithmic regularization and the incorporation of prior location-specific survey data are essential to achieve predictions that are sufficiently accurate for practical applications.

Table 3: Summary of Results: Food Security Outcomes

A. Food Consumption Score			R^2	MAE	MSE	Rank Corr.
Cross-sectional (model 1)	Spatial	OLS	<-1	11.269	Very Large	-0.004
		Post-LASSO	<-1	8.993	Very Large	0.021
		XGBOOST	undefined	9.256	165.142	undefined
Dynamic (model 2)	Temporal	OLS	0.306	6.469	103.591	0.725
		Post-LASSO	0.670	4.037	49.002	0.814
		XGBOOST	0.666	4.058	49.607	0.813
	Spatio-temporal	OLS	0.130	6.443	108.495	0.691
		Post-LASSO	0.612	4.116	49.652	0.768
		XGBOOST	0.552	4.783	59.281	0.742
Dynamic, geospatial only (3)	Spatio-temporal	OLS	<-1	12.096	311.122	0.050
		Post-LASSO	-0.166	8.726	158.572	0.073
		XGBOOST	undefined	9.728	179.058	-0.050
Dynamic, lag only (4)	Temporal	Post-LASSO	0.620	4.613	56.427	0.781
		XGBOOST	0.622	4.385	56.211	0.778
	Spatio-temporal	Post-LASSO	0.561	4.661	56.569	0.731
		XGBOOST	0.560	4.473	56.810	0.730
B. Household Hunger Scale			R^2	MAE	MSE	Rank Corr.
Cross-sectional (model 1)	Spatial	OLS	<-1	0.936	Very Large	0.051
		Post-LASSO	<-1	0.801	Very Large	0.051
		XGBOOST	undefined	0.784	1.121	undefined
Dynamic (model 2)	Temporal	OLS	0.136	0.737	1.242	0.565
		Post-LASSO	0.538	0.465	0.662	0.712
		XGBOOST	0.495	0.514	0.724	0.662
	Spatio-temporal	OLS	-0.243	0.796	1.581	0.521
		Post-LASSO	0.454	0.474	0.674	0.689
		XGBOOST	0.346	0.574	0.857	0.594
Dynamic, geospatial only (3)	Spatio-temporal	OLS	-0.794	1.035	2.316	0.084
		Post-LASSO	-0.077	0.816	1.468	0.207
		XGBOOST	undefined	0.882	1.618	0.095
Dynamic, lag only (4)	Temporal	Post-LASSO	0.426	0.523	0.821	0.602
		XGBOOST	0.416	0.555	0.836	0.594
	Spatio-temporal	Post-LASSO	0.337	0.531	0.824	0.566
		XGBOOST	0.317	0.565	0.860	0.559

Notes: For the two focus outcomes, we report the mean across all prediction folds for each metric: R2, median absolute error (MAE), mean square error (MSE) and the Rank correlation coefficient. Numbers larger than 1000 or less than -1 are marked where appropriate. The R2 and rank correlation are undefined in the case of zero-variance predictions.

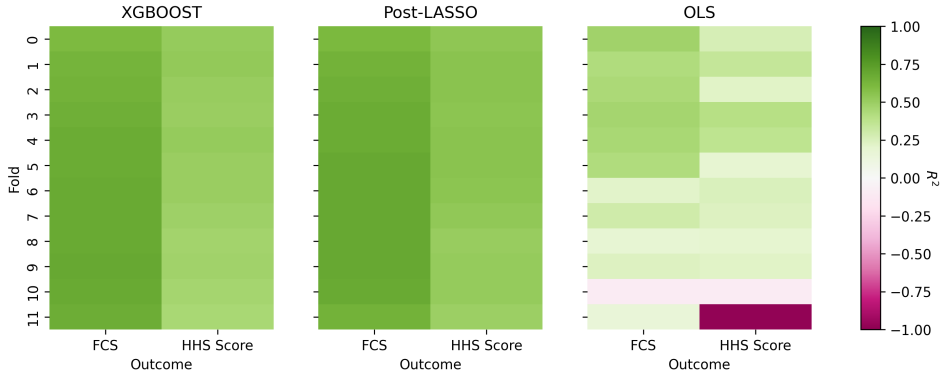
Summary results are shown in Table 3, which reports predictive performance across four metrics: the out-of-sample coefficient of determination (R^2), median absolute error (MAE), mean square error (MSE), and the rank correlation coefficient. We emphasize R^2 in the text because it is intuitive and the most commonly reported measure of predictive performance in the related literature. The R^2 is calculated relative to the mean of the observed test data; this requires the model to not only capture relative rankings between districts but also to adapt to shifts in the distribution over space and/or time. An R^2 of 1 indicates perfect prediction, a value of 0 indicates the model performs no better than predicting the mean of the *test* set, and negative values would indicate that the model performs worse than the test set mean.

Drawbacks of the R^2 are its sensitivity to variability and measurement error in the data (Corral et al., 2025), and that it does not distinguish between bias and relative ranking accuracy. In contrast, the rank correlation isolates the model's ability to correctly order communities from "best off" to "worst off," regardless of prediction bias. Another relative measure, we assess how well different quartiles of the food insecurity distribution are correctly classified in each period. Complementing these, the MAE and MSE quantify the magnitude of prediction errors in the units of the outcome, providing intuition for how far predicted food security scores deviate from reality.

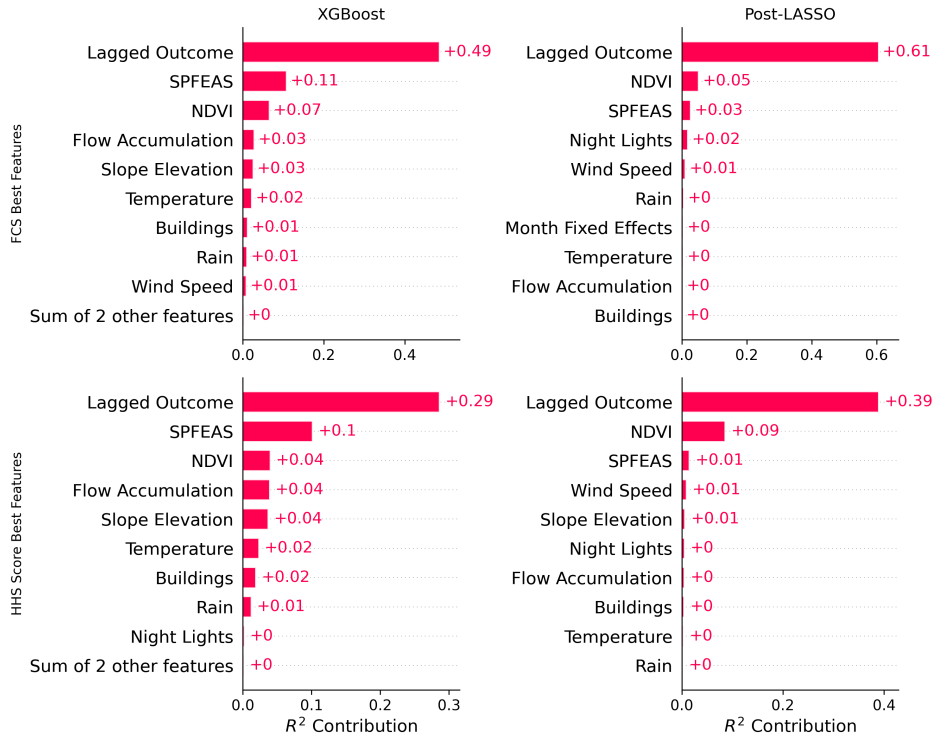
4.1.1 Temporal Prediction

Temporal cross-validation simulates prediction for populations subject to ongoing monitoring, where historical data are available for training but real-time estimates are needed to bridge the gap between survey rounds or account for data collection or processing delays. To simulate a six-month monitoring gap, we withhold the target month's survey data and predict the outcome using contemporaneous remote sensing features and the 6-month-lagged outcome value (results for 3- and 12-month lags are reported in the appendices).

We find that dietary diversity (FCS) is the most predictable outcome, with both XGBoost



(a) R^2 Heatmaps of Time Models



(b) Shapley-R^2 Values of Time Models for XGBoost & Post-LASSO for FCS & HHS Score

Figure 3: Performance of Time Models

Notes: Figures show results of Time models. In Panel (a), colors range from red to green; redder corresponding to negative R^2 and greener corresponding to higher R^2. Colors correspond to the mean R^2 across all folds in the month-year observations. Results shown are the average scores across block cross-validation corresponding to Figure 2. As illustrated in Figure 2, higher folds use older (sometimes less) training data and are used to predict further forward in time. Panel (b) shows bar plots of the absolute value of Shapley-decomposed R^2 values for XGBoost and Post-LASSO methods averaged across all time cross-validation variants. Higher values represent more importance. Shapley values were calculated using the `shap` library in Python using a tree explanation with XGBoost and linear explanation for Post-LASSO. Shapley values for each feature group are the sum of Shapley values for each feature in the group. The SPFEAS group includes all SPFEAS features, excluding NDVI, which overlaps with the interpretable feature set. R^2 decomposition was calculated based on Redell (2019).

and Post-LASSO achieving an average R^2 of 0.67. Predictive accuracy for the HHS is slightly lower, with R^2 values of 0.54 for Post-LASSO and 0.50 for XGBoost. While simple OLS models achieve positive R^2 values, they tend to perform poorly in higher folds, when more periods of testing data are predicted. Panel (a) of Figure 3 displays variation in the average R^2 for these models across the temporal cross-validation folds.

Additional metrics reinforce and clarify these findings. Our XGBoost and Post-LASSO models achieve MAEs equivalent to approximately one-third to one-half of the typical standard deviation for each outcome.[3] These models achieve rank correlations of 0.81 for FCS and 0.66–0.71 for HHS, indicating a moderate to strong ability to correctly distinguish relative levels of food insecurity.

To visualize how well these models identify the most vulnerable communities in any given period, we conduct a quartile analysis comparing true observed versus predicted food insecurity quartiles. Enumeration areas (EAs) in the bottom quartile for dietary diversity (FCS) are consistently identified with high accuracy (83% mean, 8% SD across periods). While classification accuracy is lower and more variable for household hunger (HHS) (50% mean, 20% SD), both models substantially outperform the 25% baseline expected from random classification.

The relatively strong model performance in predicting food insecurity over time highlights the potential of these methods for food insecurity monitoring. However, model performance depends heavily on the inclusion of the lagged outcome. When we estimate a "geospatial-only" temporal model (equation 3), predictive performance declines precipitously. We explore this pattern further in Section 4.1.4 through an analysis of feature importance.

While temporal prediction is most relevant for early warning in areas with established data monitoring systems, humanitarian decision-makers frequently face the challenge of targeting resources to locations where historical data are unavailable or more limited. To address this, we next evaluate the models' ability to extrapolate to new locations under

[3]See Online Appendix Table S2 for descriptive statistics, including the quarterly mean and standard deviation of FCS and HHS.

two constraints: predicting into districts with no survey data at all ("spatial models") and predicting into districts excluded from training, where historical data is limited to a single lagged outcome ("spatio-temporal models").

4.1.2 Spatial Prediction

The cross-sectional "spatial models" simulate a scenario where the practitioner has a single cross-sectional survey of about 150 communities and wants to extrapolate outcomes for the same period to communities in neighboring districts. The data is subset to a particular month-year and the model evaluated using LODO cross-validation. In this setting, we find that predicting into other districts using purely geospatial data is not feasible for these food security outcomes. Using OLS, Post-LASSO, or XGBoost, there is little signal between geospatial features and measures of food security. The XGBoost algorithm often breaks down and predicts the mean value of the training set, resulting in zero variance prediction across clusters.[4] The heatmap of R^2 values by fold for spatial prediction is shown in Online Appendix Figure S8.

4.1.3 Spatio-Temporal Prediction

Next, we consider the "spatio-temporal" models. This involves constraining the training data so that the model has not 'seen' any data from the predicted district or the predicted time period. This exercise simulates real constraints that are faced by development organizations: ongoing data collection may occur in some districts (e.g., the RFMS collects data in 10 of Malawi's 28 districts), while others will not be included in the survey but could have lagged outcome data available from another data source.

Compared to the temporal models, the Post-LASSO model performance only decreases slightly when trained on out-of-district data, to an average R^2 of 0.61 for FCS and 0.45 for HHS. The XGBoost is more sensitive to the change, with an R^2 of about 0.55 for FCS and

[4] In such cases the R^2 and rank correlation coefficients are undefined. In summary statistics and visualizations, models with zero variance prediction are assigned an $R^2 = 0$.

0.35 for HHS. This pattern is consistent across the additional performance metrics (MAE, MSE, and rank correlation); the Post-LASSO models generally maintain similar performance to the temporal case, while XGBoost performance noticeably decreases. This suggests that the XGBoost models overfit to the specific context of the training districts. Consistent with this, learning curves reveal a persistent gap between training and testing performance for XGBoost models, even at the full sample size, whereas Post-LASSO models exhibit smaller gaps and approach convergence (Online Appendix Figure S25).

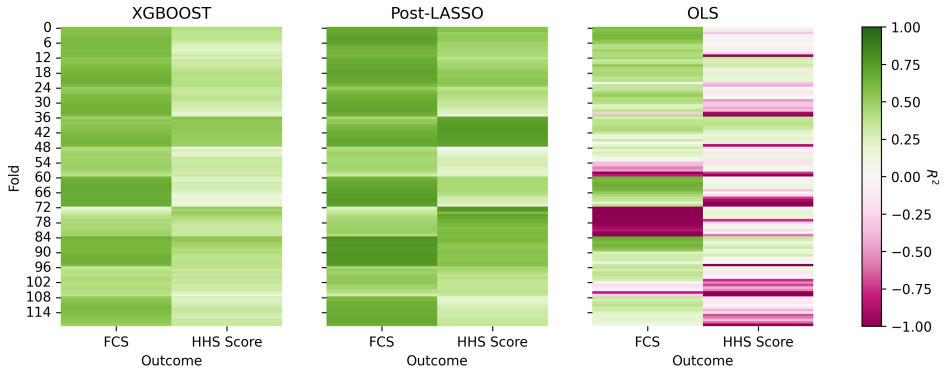
Variation in performance across cross-validation folds is depicted in Panel (a) of Figure 4. The structure of these heatmaps again corresponds to Panel (a) of Figure 3, but with the added district dimension: for each held out district, we test the full set of cross-validation folds as in the "temporal" cross-validation.

4.1.4 Feature Importance for Predicting Food Insecurity

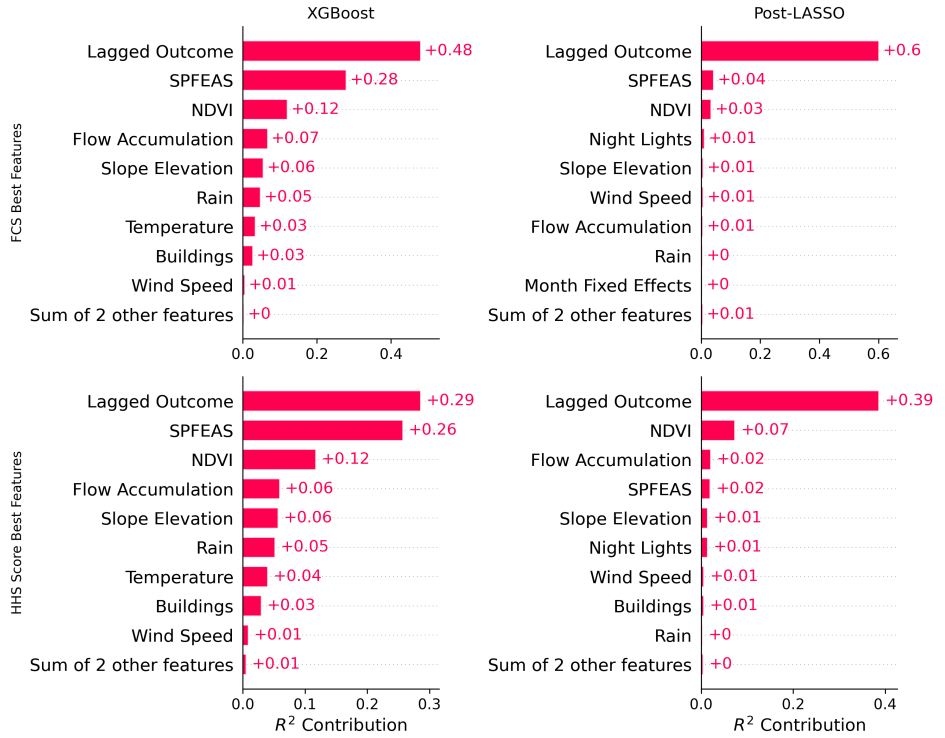
To understand the drivers of model performance, we first consider the feature importance for our highest-performing "temporal" models. Panel (b) of Figure 3 presents feature contributions calculated using Shapley-decomposed R^2 values.[5] Importantly, Shapley values explain the behavior of the model rather than the true underlying data-generating process, and must be interpreted with caution. Consequently, we focus our analysis on FCS and HHS; because illness and schooling outcomes were not successfully predicted, their Shapley values would largely reflect the model fitting to statistical noise (Section 4.2).

Feature importance varies by outcome and model, but in all cases the most important feature is the lagged outcome. Both our interpretable features (i.e., vegetation, weather, and geography) and contextual features (the SpFeas excluding NDVI, which is considered separately) contribute to prediction, with the interpretable features contributing more overall

[5]Shapley values provide a way to gauge the importance of a feature used in prediction, by treating each feature as a player in a cooperative game and calculating their share of the gain in a coalition, where the "gain" is the explained variance of the model. The units of these Shapley values are in the units of the outcome of interest. We use a Shapley value variance decomposition and re-calculate these values as per-feature R^2 contributions to the prediction (Redell, 2019).



(a) R^2 Heatmaps of Space-Time Models



(b) Shapley-R^2 Values of Space-Time Models, XGBoost & Post-LASSO for FCS & HHS Score

Figure 4: Performance of Space-Time Models

Notes: Figures show results of Space-Time models. In Panel (a), colors range from red to green; redder corresponding to negative R^2 and greener corresponding to higher R^2. Colors correspond to the mean R^2 across all folds in the district-month-year observations. Results shown are the average scores across block cross-validation (Figure 2), but further divided to display variation by the test district within each temporal fold. As illustrated in Figure 2, higher folds use older (sometimes less) training data and are used to predict further forward in time. Panel (b) shows bar plots of the absolute value of Shapley-decomposed R^2 values for XGBoost and Post-LASSO methods averaged across all space-time cross-validation variants. Higher values represent more importance. Shapley values were calculated using the `shap` library in Python using a tree explanation with XGBoost and linear explanation for Post-LASSO. The SPFEAS group includes all SPFEAS features, excluding NDVI, which overlaps with the interpretable feature set. Shapley values for each feature group are the sum of Shapley values for each feature in the group. R^2 decomposition was calculated based on Redell (2019).

across all models. Appendix Figures S26–S29 report the stability of feature importance. While there is some variability across folds, the leading features—lagged outcomes, the contextual feature group, and NDVI—consistently contribute to prediction across models and folds.[6]

For the FCS XGBoost model, NDVI features are the third most important feature; they rank second for Post-LASSO HHS Score. NDVI is correlated (albeit imperfectly) to biomass growth and, therefore, may capture seasonality and/or inter-annual variation in food supply. Time variant weather variables, including levels and anomalies in rainfall, temperature, and cyclone winds, contribute to prediction – particularly in the Post-LASSO models. Features that identify buildings are sometimes important, particularly for FCS, which may point to the relationship between food security, building size and density as a proxy for household wealth or income. For HHS, slope, elevation, and flow accumulation are relatively important predictors, which are related to agricultural productivity as well as an area's potential vulnerability in the case of a shock. A greater flow accumulation can result in standing water and flooding during heavy rainfall, which may increase household vulnerability to shock events. On the other hand, leveraging residual moisture in floodplains for planting may be an important secondary source of food, especially in years when rains are insufficient or unpredictable during the primary growing season.

Time invariant features are overall less important, but contribute more to the XGBoost models. This may be because XGBoost creates nonlinear interactions to boost performance. The Post-LASSO, however, did not include interactions with other features in order test the predictive power of each feature separately.

Shapley values are not computed for the poorly performing "spatial" models. Results for the "spatio-temporal" models are shown in Panel (b) of Figure 4. Despite these models needing predictive power both from the time-series and cross-sectional dimension, they look broadly similar to the Shapley graphs in Figure 3. However, spatial features such as

[6]The exception is when the Post-LASSO model does not select contextual features (Figures S26 and S27).

the SpFeas and NDVI features are more important to the predictions, particularly in the XGBoost models.

4.2 Nowcasting Illness and Schooling Disruptions

In contrast to food insecurity, our models fail to reliably predict illness and schooling disruptions (Appendix Tables S6 and S7 report summary results). XGBoost models consistently collapse to the mean of the training data, resulting in undefined R^2 and rank correlation coefficients. While Post-LASSO specifications outperform OLS, they fall short of the accuracy and reliability needed for practical application in the temporal, spatial, or spatio-temporal scenarios tested. For the "temporal" models the average R^2 for the Post-LASSO models was 0.08 for illness, 0.11 for illness under five, and 0.07 for schooling disruptions. In the "spatio-temporal" case the average R^2 are slightly negative; in the "spatial" case, even more-so.

While predictive accuracy is poor in terms of explained variance, the Post-LASSO illness and child illness models achieve moderate rank correlation coefficients of 0.32 to 0.49. At first glance, this suggests the model confers some ability to distinguish relative disease burden. However, when compared to a lag-only Post-LASSO baseline, we find that both R^2 and rank correlation increase when geospatial features are removed. This suggests that the models fail to capture meaningful signal from the remote sensing features, yet it highlights that the high-frequency survey data may still have inherent value for monitoring such outcomes.

As we explore further in the Discussion, we attribute the failure of geospatial features to enhance prediction for these outcomes to three likely factors. First, measurement error may be significant; unlike the well-validated food security indices, our illness and schooling measures are simple self-reports that have not been validated against more robust indicators of disease burden. Second, data limitations may matter. Given the stochastic nature of disease outbreaks, predictions may be particularly sensitive to the modest sample size. Third, these outcomes may suffer from a weak structural link to geospatial features. For example, school attendance is ultimately a household decision. While these choices may be influenced

by detectable phenomena like drought or flooding, the relationship is not deterministic and correlations may therefore be weak.

Finally, we consider whether our overall illness outcome may obscure more predictable patterns in specific types of illness. While the self-reported nature of the data prohibits us from identifying specific pathogens, respondents do distinguish between different symptoms. Testing these narrower definitions, we find that the performance of the Post-LASSO model increases very slightly, from an R^2 of 0.08 for overall illness to 0.12 for fever, 0.15 for respiratory illness, and 0.15 for digestive illness (Appendix Table S8). Because broad symptom categories may obscure the environmental drivers of specific diseases, and because self-reports introduce measurement error, future models trained on verified clinical data for specific pathogens might yield better predictions.

Similarly, modeling schooling disruptions may benefit from the incorporation of administrative records on enrollment, attendance, or school operating status. Furthermore, because shock-induced school absences are relatively rare events in our sample, algorithms may struggle to learn to predict these events. In such cases of severe class imbalance, specialized data sampling techniques might offer modest predictive gains (Zhou et al., 2022).

4.3 Heterogeneity and Robustness

Given the dependence on lagged outcomes for predictions, it is particularly important to consider heterogeneity in predictive performance. Even if the average R^2 is high enough to predict average trends for a specific district, performance may be uneven across the year, especially when an adverse shock renders past food security a less useful predictor of current conditions. Online Appendix Figures S17-S20 plot period-specific performance (the main results are summary statistics at the fold-level). We find that while model performance does vary by period, consistency is improved by the addition of geospatial variables. To visualize this, one can compare the full models (equation 2) in Online Appendix Section 8.9 to the equivalent lag-only models (equation 4) in Online Appendix Section 8.10.

We also plot realized versus average predicted values of FCS and HHS across time periods (Appendix Figure S16). We find that training models on data from the first year of collection, when COVID-19-related policies were most likely to introduce structural differences, does not substantially affect model performance and we do not observe a consistent seasonal pattern in predictive accuracy.[7] While modeled predictions broadly capture the average dynamics of food insecurity in the study area, both Post-LASSO and XGBoost models on average over-predict dietary diversity during the 2023 post-harvest (May-October) season and under-predict a spike in HHS in the late fall of 2023 (October-November). This period was characterized by high maize prices following an exceptionally weak harvest in southern Malawi. This prediction bias serves as a critical caution for early warning applications: severe shocks that lack common support in the model's historical training are likely to result in an under-estimation of acute food insecurity.

Our primary models use a 6-month lag, but we also assess prediction 3 to 12 months forward in time from the lagged survey observation. We find that performance degrades slightly as this temporal lag increases. For temporal predictions of dietary diversity (FCS), XGBoost (Post-LASSO) models achieve average R^2 values of about 0.70 (0.73) at 3 months, 0.67 (0.67) at 6 months and 0.54 (0.61) at 12 months. The pattern is similar for HHS adverse coping strategies, but with lower overall R^2 values of 0.57 (0.56) at 3 months, 0.49 (0.54) at 6 months and 0.38 (0.40) at 12 months.

While Shapley values are a useful tool for comparing the relative contributions of different variables to our nowcasts, we interpret them with caution due to the limited accuracy of those predictions. As a robustness check on the value addition of geospatial features, we re-estimate models excluding the lagged outcome data (Equation 3) as well as only lagged outcome data (Equation 4) using spatio-temporal cross-validation. Summary results are reported in Table 3 for food security outcomes (in Appendix Tables S6 and S7 and for illness and schooling).

The lag-only results confirm that the addition of the geo-spatial features modestly im-

[7]However, our testing set spans too short a period to confidently assess seasonal variation in predictive performance and we cannot rule out a COVID-19 bias.

proves the accuracy of nowcasts. This is most pronounced for the Post-LASSO models–our strongest predictors. Without the geospatial features, our R^2 decreases by approximately 5 percentage points for FCS and 11 to 12 percentage points for HHS. For XGBoost, accuracy generally improves with the addition of the geospatial features to the models; the exception is the spatio-temporal FCS models, which explain a similar share of variation with and without the geospatial features.

Consistent with the results from the cross-sectional models, the purely geospatial spatio-temporal models do not reliably predict any of our five outcome variables; the R^2 values are frequently negative and the XGBoost models often break down. The heat maps for the no-lag results are plotted in Appendix Figure S11.

Data reduction might mitigate the overfitting we observe, particularly in the case of the XGBoost food insecurity models. So we apply principal component analysis (PCA) to the contextual features to test this approach. However, it does not meaningfully improve XGBoost or Post-LASSO performance. OLS models improve, but still perform poorly overall. These PCA results are reported in Online Appendix (Tables S12 and S13).

In sum, predicting the time series of unsampled districts becomes significantly more achievable when there is some prior outcome information available for those unsampled districts. This need not be from high-frequency data and need not be current. However, having some lagged data is crucial, even with the inclusion of remote sensing and geographic information.

5 Discussion

Enhanced real-time monitoring of food insecurity, health, and education can guide timely interventions to alleviate unnecessary suffering and prevent lasting damage to human capital. To this end, we design and test an approach to nowcast outcomes that may signal disruption—or even reversal—of human capital accumulation. This builds on a litera-

ture highlighting how short-term shocks can have lasting impacts, especially in vulnerable low-income, rain-fed agricultural settings where coping resources are limited across households, markets, and the public sector (Hoddinott and Kinsey, 2001; Alderman et al., 2006; Björkman-Nyqvist, 2013; Blom et al., 2022).

Our approach leverages temporal autocorrelation from high-frequency household survey data, together with indicators of causal drivers, real-time outcome proxies, and (less interpretable) contextual features, all integrated at a monthly time step within flexible machine-learning models. The results underscore both the potential and the limitations of these methods when combining monthly household panel data with geospatial features to nowcast human capital outcomes.

First, we find that machine learning models can successfully predict community-level food insecurity forward in time and into neighboring districts, provided that at least one survey round from up to twelve months prior is available. These models are trained on as little as one year of historical high-frequency data. Second, remotely sensed feature sets are not well-suited for stand-alone prediction across space in this context. However, when integrated with lagged survey data, geospatial features meaningfully improve both the overall accuracy and the consistency of food security nowcasts. In Post-LASSO models, which on average explain the most variation in our measures of food security, the addition of geospatial features increases the R^2 value of nowcasts by approximately 0.05 for dietary diversity (FCS) and 0.11 to 0.12 for household hunger (HHS). Third, in contrast to food insecurity, this data fusion approach fails to reliably predict measures of self-reported illnesses and schooling disruptions. This is likely due to a combination of measurement error, the stochastic nature of these variables, and the fact that idiosyncratic household decisions lack strong, deterministic links to geospatial features. Finally, methodologically, we find that regularized linear models (Post-LASSO) generally match or slightly outperform the tree-based XGBoost algorithm, which tends to overfit at the sample size (178 village clusters) of our high-frequency household survey.

5.1 Key Challenges and Limitations

Data fusion has increasingly diverse policy applications, ranging from understanding long-standing structural inequalities of wealth and development to anticipating and responding to rapidly unfolding crises. Not all tasks will be equally served by a single set of training data, algorithms, or project structure. Nowcasting shocks to human capital formation differs fundamentally from predicting spatial variation in stable stocks of asset wealth or health, knowledge, skills, and abilities, which accumulate over lifetimes and may require different data and methods. Consequently, predicting these dynamic outcomes presents unique methodological and practical hurdles. While the precise drivers of variable model performance are complex, several structural factors may limit predictive accuracy in this context and inform directions for future research.

First, human capital shocks at the village level are difficult to detect from satellite imagery. Even if we can proxy the geography, built environment, and the timing and quality of local harvests through weather and vegetation indicators, many important structural, time-varying determinants of food insecurity remain missing. The availability and price of food and other essential goods, local labor market conditions, monetary policy, and social safety nets stand out as important omitted drivers of food insecurity poorly captured by remotely sensed features. Similarly, we lacked data on the infrastructure that supplies human capital formation (e.g., schools, health centers). Finally, households' coping strategies during periods of scarcity are heterogeneous—especially regarding decisions about health and schooling—meaning that even a perfectly characterized environmental shock cannot be easily mapped to outcomes like illness and disrupted schooling.

Second, the training data in this case may be insufficient to effectively identify key relationships between the geospatial features and outcome variables, at least for tree-based models. Our training data consist of $n = 178$ clusters over approximately $t = 12$ or $t = 24$ time periods, yielding over two thousand observations, and in some model variants, over four thousand. Each cluster contains an average of approximately 25 households. Although this

would be large for many regression applications, it is small for machine learning, and the panel structure means the observations are not independent, reducing the effective sample size. This may explain why Post-LASSO outperforms XGBoost in most models. Although XGBoost is adept at learning complex relationships from a large and potentially poorly specified feature set, and may therefore have a higher upper bound on predictive performance, it typically requires large training datasets to do this well. In contrast, Post-LASSO results appear reasonably stable, with the test–train performance gap increasing only modestly when the training sample is reduced by up to 25% (Online Appendix Figure S25). Prediction using a strictly regularized linear model may be viable even with smaller sample sizes, whereas tree-based models may be more likely to improve with additional training data.

Relatedly, measurement error in the training data may limit model performance, particularly for the illness and schooling disruption outcomes. While the FCS (for dietary diversity) and HHS (for adverse coping strategies) are well-validated measures of food (in)security, measured using established techniques, the illness and schooling outcomes are simple self-reports. We lack evidence on whether self-reports of illness accurately reflect disease burden or how reliably individuals report these outcomes.

Third, our study area of southern Malawi is unusual in two important ways. One is that the sample is relatively homogeneous; RFMS districts were selected specifically based on their vulnerability and our sample is entirely rural. Often, the overall predictive performance in geospatial modeling of welfare outcomes relies on the ability of these models to distinguish rural versus urban areas.[8] Food systems and the supply of human capital services differ markedly in rural versus urban areas. Another important contextual factor is that Malawi experiences heavy cloud cover during its rainy season. This introduces noise into publicly available indicators derived from imagery, as well as the spatial features calculated here. Using remote sensing in other contexts may yield better results.

Fourth, a general caveat to these results is that while a high R^2 may indicate that the

[8]Predicting asset wealth, Yeh et al. (2020) achieve an overall R^2 of about 0.70, while their R^2 for rural areas only is 0.32.

models can predict average trends for a specific district, they may not be suitable for forecasting anomalous or novel events, such as environmental disasters or other shocks that did not occur during the training period. Although our sample includes periods when enumeration areas endured shocks, they may not have occurred with enough variation and frequency for the models to "learn" how geospatial indicators reflect these shocks——a challenge exacerbated by the limited sample size here and typical of household surveys. Our results suggest that these models will tend to under-predict outcomes of concern during shock events that generate strong downturns in dietary diversity and household hunger.

5.2 Future Research and Policy Implications

Additional training data could plausibly improve predictive accuracy (Gualavisi and Newhouse, 2025; Zheng et al., 2025), as well as our ability to understand uncertainty and interpret results. Future research could usefully investigate the extent to which increasing the number of villages or the duration of the panel, to learn from more varied constellations of growing conditions and other shocks, or utilizing a more heterogeneous sample of villages would enhance model performance and reduce the risk of overfitting.

With a limited training sample, carefully crafted feature sets informed by domain knowledge—perhaps including data on human capital services like school and health facility maps—could improve outcomes. Indeed, given that contextual features are computationally intensive, lack direct interpretability, and ultimately offered a lower predictive contribution in our models, interpretable features may be the more practical pathway forward in the near term. Methodologically, training deep neural networks from scratch is likely to be infeasible given the scale of typical household survey datasets. However, transfer learning and the use of geospatial foundation models may offer a promising workaround to this data constraint (Zheng et al., 2025). Whether such approaches can succeed at the scale of our current sample or similar monitoring systems remains an important open question for future work.

This paper highlights the value of high-frequency household survey data as both a moni-

toring tool and to train nowcasting models. A system like the RFMS enables data availability with approximately a one-month lag, significantly faster than the typical 1–2 year delay associated with traditional household living standards surveys, making it possible to generate nowcasts in near-real time. As national statistical institutes and other data collection agencies consider their portfolios, our findings offer evidence on the potential usefulness of such data. Phone-based surveys may also be an effective source of training data. These became more common during the COVID-19 pandemic (Gourlay et al., 2021), and can be deployed where in-person surveys are unsafe, infeasible, or cost-prohibitive. Ongoing work aims to understand and correct for potential sampling biases in phone-based surveys (Dillon et al., 2025).

While our study demonstrates the value of high-frequency surveys, it also highlights the limitations of relying on simple self-reports, particularly when measuring illness and schooling disruptions. Harnessing alternative data streams—such as verified records of specific pathogens, school enrollment, daily attendance, and school operating status—could offer a highly productive avenue for future data fusion efforts. Working with policymakers to construct training datasets around precise, operationally relevant outcomes that measure shocks to human capital, and evaluating model performance based on the real-world inclusion and exclusion errors that they care about, represents a critical next step for this line of research.

Expanding the study area to multiple countries and both urban and rural populations, along with larger training datasets, could improve our understanding of spatial and temporal extrapolation. This is important because, even with methodological advances and investment, high-frequency survey coverage is likely to remain limited. Focusing specifically on predicting the impact of shocks would be another useful area for future research. Our sample is geographically and temporally limited, but it is encouraging that we do not observe evidence of degradation in model performance as the gap between training and testing periods increases. Moreover, performance declines only modestly when predicting into neighboring districts where we simulate that only a single round of lagged data is available.

To design and target development and humanitarian assistance, policymakers require information about structural well-being and the shocks that threaten to erode it. Here, we focus on the latter: our ability to nowcast periods of food insecurity, ill health, and disrupted schooling, all of which are indicative of acute scarcity and disinvestment in human capital. From a data fusion perspective, this is a fundamentally different task than downscaling estimates of asset wealth or consumption expenditures. Our results suggest that publicly available geospatial features can meaningfully improve the accuracy of nowcasts for dietary diversity and household hunger. At the same time, household surveys remain the backbone of humanitarian monitoring systems. Maintaining large-sample and geographically heterogeneous household surveys is crucial to fully exploiting the potential of geospatial data for early warning.

Data Availability Statement

The data underlying this article cannot be shared publicly for the privacy of individuals that participated in the study. All code is publicly available at: [DOI to be included in final version].

Author Contribution Statement

E.T. and A.M. contributed equally to this research. *Conceptualization*: C.B., A.M., D.N., E.T., J.U., M.W; *Methodology*: A.M., D.N., E.T.; *Data curation*: E.T. (lead), M.M. (SpFeas lead), J.U. (RFMS lead), A.C., R.E.; *Formal analysis*: A.M. (lead), E.T.; *Data visualization*: A.M. (lead), M.M., E.T., J.U.; *Writing – original draft*: E.T. (lead), A.M., M.M., J.U.; *Writing – review & editing*: E.T. (lead), C.B., M.M., A.M., D.N., J.U., M.W.; *Funding acquisition & project administration*: C.B., A.M., D.N., E.T., J.U., M.W.

References

Abate, G. T., de Brauw, A., Hirvonen, K., and Wolle, A. (2023). Measuring consumption over the phone: Evidence from a survey experiment in urban Ethiopia. *Journal of Development Economics*, 161:103026.

Abay, K. A., Berhane, G., Hoddinott, J., and Tafere, K. (2022). Respondent Fatigue Reduces Dietary Diversity Scores Reported from Mobile Phone Surveys in Ethiopia during the COVID-19 Pandemic. *The Journal of Nutrition*, 152(10):2269–2276.

Acemoglu, D., Gallego, F. A., and Robinson, J. A. (2014). Institutions, Human Capital, and Development*. *Annual Review of Economics*, 6(Volume 6, 2014):875–912.

Adair, L. S., Fall, C. H., Osmond, C., Stein, A. D., Martorell, R., Ramirez-Zea, M., Sachdev, H. S., Dahly, D. L., Bas, I., Norris, S. A., Micklesfield, L., Hallal, P., and Victora, C. G. (2013). Associations of linear growth and relative weight gain during early life with adult health and human capital in countries of low and middle income: Findings from five birth cohort studies. *The Lancet*, 382(9891):525–534.

Agamile, P. and Lawson, D. (2021). Rainfall shocks and children's school attendance: Evidence from Uganda. *Oxford Development Studies*, 49(3):291–309.

Aiken, E., Bellue, S., Karlan, D., Udry, C., and Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903):864–870.

Alderman, H., Hoddinott, J., and Kinsey, B. (2006). Long Term Consequences of Early Childhood Malnutrition. *Oxford Economic Papers*, 58(3):450–474.

Anderson, C. L., Reynolds, T., Merfeld, J. D., and Biscaye, P. (2018). Relating Seasonal Hunger and Prevention and Coping Strategies: A Panel Analysis of Malawian Farm Households. *Journal of Development Studies*, 54(10):1737–1755.

Andree, B. P. J., Chamorro, A., Kraay, A., Spencer, P., and Wang, D. (2020). *Predicting food crises*. The World Bank.

Anttila-Hughes, J. and Hsiang, S. (2013). Destruction, Disinvestment, and Death: Economic and Human Losses Following Environmental Disaster. SSRN Scholarly Paper ID 2220501, Social Science Research Network, Rochester, NY.

Anttila-Hughes, J. K., Jina, A. S., and McCord, G. C. (2021). ENSO impacts child undernutrition in the global tropics. *Nature Communications*, 12(1):5785.

Ballard, T., Coates, J., Swindale, A., and Deitchler, M. (2011). Household Hunger Scale: Indicator Definition and Measurement Guide. Technical Report Food and Nutrition Technical Assistance II Project, FHI 360., USAID, Washington, D.C.

Barrett, C. B. (2025). Second best underestimates of malnutrition in an era of multiplying food crises. *Nature Food*, 6(12):1109–1110.

Beegle, K., Dehejia, R. H., and Gatti, R. (2006). Child labor and agricultural shocks. *Journal of Development Economics*, 81(1):80–96.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Björkman-Nyqvist, M. (2013). Income shocks and gender gaps in education: Evidence from Uganda. *Journal of Development Economics*, 105:237–253.

Blom, S., Ortiz-Bobea, A., and Hoddinott, J. (2022). Heat exposure and child nutrition: Evidence from West Africa. *Journal of Environmental Economics and Management*, 115:102698.

Browne, C., Matteson, D. S., McBride, L., Hu, L., Liu, Y., Sun, Y., Wen, J., and Barrett, C. B. (2021). Multivariate random forest prediction of poverty and malnutrition prevalence. *PLOS ONE*, 16(9):e0255519.

Burke, M., Driscoll, A., Lobell, D. B., and Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535).

Chao, S., Engstrom, R., Mann, M., and Bedada, A. (2021). Evaluating the Ability to Use Contextual Features Derived from Multi-Scale Satellite Imagery to Map Spatial Patterns of Urban Attributes and Population Distributions. *Remote Sensing*, 13(19):3962.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Constenla-Villoslada, S., Liu, Y., McBride, L., Ouma, C., Mutanda, N., and Barrett, C. B. (2025). High-frequency monitoring enables machine learning–based forecasting of acute child malnutrition for early warning. *Proceedings of the National Academy of Sciences*, 122(23):e2416161122.

Corral, P., Henderson, H., and Segovia, S. (2025). Poverty mapping in the age of machine learning. *Journal of Development Economics*, 172:103377.

D. Matarira, O. M. and Naidu, M. (2022). Texture analysis approaches in modelling informal settlements: a review. *Geocarto International*, 37(26):13451–13478.

Dillon, B., Twivwe, S., and Upton, J. (2025). Phone Tree Surveys and the Wisdom of Crowds. Working Paper.

Duque, J. C., Patino, J. E., Ruiz, L. A., and Pardo-Pascual, J. E. (2015). Measuring intraurban poverty using land cover and texture metrics derived from remote sensing data. *Landscape and Urban Planning*, 135:11–21.

Ellis, F. and Manda, E. (2012). Seasonal Food Crises and Policy Responses: A Narrative Account of Three Food Security Crises in Malawi. *World Development*, 40(7):1407–1417.

Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B., and Bright, E. (2009). A global poverty map derived from satellite data. *Computers & Geosciences*, 35(8):1652–1660.

Engstrom, R., Sandborn, A., Qin, Y., Burgdorfer, J., Stow, D., Weeks, J., and Graesser, J. (2015). Mapping slums using spatial features in Accra, Ghana. In *2015 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. IEEE.

Fishman, R., Carrillo, P., and Russ, J. (2019). Long-term impacts of exposure to high temperatures on human capital and economic productivity. *Journal of Environmental Economics and Management*, 93:221–238.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378.

Gilbert, C., Christiaensen, L., and Kaminski, J. (2017). Food price seasonality in Africa: Measurement and extent. *Food Policy*, 67:119–132.

Glick, P. J., Sahn, D. E., and Walker, T. F. (2016). Household Shocks and Education Investments in Madagascar. *Oxford Bulletin of Economics and Statistics*, 78(6):792–813.

Gourlay, S., Kilic, T., Martuscelli, A., Wollburg, P., and Zezza, A. (2021). Viewpoint: High-frequency phone surveys on COVID-19: Good practices, open questions. *Food Policy*, 105:102153.

Graesser, J., Cheriyadat, A., Vatsavai, R. R., Chandola, V., Long, J., and Bright, E. (2012). Image based characterization of formal and informal neighborhoods in an urban landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(4):1164–1176.

Gualavisi, M. and Newhouse, D. (2025). Integrating survey and geospatial data for geographical targeting of the poor and vulnerable: Evidence from Malawi. *The World Bank Economic Review*, 39(2):377–409.

Hajison, P. L., Mwakikunga, B. W., Mathanga, D. P., and Feresu, S. A. (2017). Seasonal variation of malaria cases in children aged less than 5 years old following weather change

in Zomba district, Malawi. *Malaria Journal*, 16(1):264.

Head, A., Manguin, M., Tran, N., and Blumenstock, J. E. (2017). Can human development be measured with satellite imagery? *Proceedings of ICTD '17, Lahore, Pakistan, November 16–19, 2017*.

Hoddinott, J. and Kinsey, B. (2001). Child Growth in the Time of Drought. *Oxford Bulletin of Economics and Statistics*, 63(4):409–436.

ILO (2026). ILO modeled estimates: Methodological overview.

Kirolos, A., Thindwa, D., Khundi, M., Burke, R. M., Henrion, M. Y. R., Nakamura, I., Divala, T. H., Nliwasa, M., Corbett, E. L., and MacPherson, P. (2021). Tuberculosis case notifications in Malawi have strong seasonal and weather-related trends. *Scientific Reports*, 11:4621.

Lentz, E., Baylis, K., Michelson, H., and Kim, C. (2025). Official estimates of global food insecurity undercount acute hunger. *Nature Food*, 6(12):1196–1208.

Lentz, E. C. and Maxwell, D. (2022). How do information problems constrain anticipating, mitigating, and responding to crises? *International Journal of Disaster Risk Reduction*, 81:103242.

Lentz, E. C., Michelson, H., Baylis, K., and Zhou, Y. (2019). A data-driven approach improves food insecurity crisis prediction. *World Development*, 122:399–409.

Maccini, S. and Yang, D. (2009). Under the Weather: Health, Schooling, and Economic Consequences of Early-Life Rainfall. *American Economic Review*, 99(3):1006–1026.

Mahler, D. G., Castañeda Aguilar, R. A., and Newhouse, D. (2022). Nowcasting Global Poverty. *The World Bank Economic Review*, page lhac017.

Marty, R. and Duhaut, A. (2024). Global poverty estimation using private and public sector big data sources. *Scientific Reports*, 14(1):3160.

Maxwell, D. and Hailey, P. (2021). Analysing Famine: The Politics of Information and Analysis in Food Security Crises. *Journal of Humanitarian Affairs*, 3(1):16–27.

McBride, L., Barrett, C. B., Browne, C., Hu, L., Liu, Y., Matteson, D. S., Sun, Y., and Wen, J. (2021). Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning. *Applied Economic Perspectives and Policy*, 44(3):879–892.

Miggo, M., Harawa, G., Kangwerema, A., Knovicks, S., Mfune, C., Safari, J., Kaunda, J. T., Kalua, J., Sefu, G., Phiri, E., and Patel, P. (2023). Fight against cholera outbreak, efforts and challenges in Malawi. *Health Science Reports*, 6(10):e1594.

Mude, A. G., Barrett, C. B., McPeak, J. G., Kaitho, R., and Kristjanson, P. (2009). Empirical forecasting of slow-onset disasters for improved emergency response: An application to Kenya's arid north. *Food Policy*, 34(4):329–339.

Newhouse, D. (2024). Small area estimation of poverty and wealth using geospatial data: What have we learned so far? *Calcutta Statistical Association Bulletin*, page 00080683231198591.

Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica*, 49(6):1417–1426.

Noor, A. M., Alegana, V. A., Gething, P. W., Tatem, A. J., and Snow, R. W. (2008). Using remotely sensed night-time light as a proxy for poverty in Africa. *Population Health Metrics*, 6(1):5.

NSO (2020). The Fifth Integrated Household Survey (IHS5) 2020 Report. Technical report, National Statistical Office of Malawi.

Pandey, P., Kington, J., Kanwar, A., Simmon, R., and Abraham, L. (2023). Planet Basemaps for NICFI Data Program. Addendum to Basemaps Product Specification.

Redell, N. (2019). Shapley decomposition of r-squared in machine learning models. *arXiv preprint arXiv:1908.09718*.

Rossi, F. (2020). Human Capital and Macroeconomic Development: A Review of the Evidence. *The World Bank Research Observer*, 35(2):227–262.

Rossi, F. and Weber, M. (2024). The Accumulation and Utilization of Human Capital over the Development Spectrum. *World Bank Policy Research Working Paper*, (WPS10891).

Staffieri, I., Sitko, N. J., and Maluccio, J. A. (2023). Sustaining enrolment when rains fail: School feeding, rainfall shocks and schooling in Malawi. *Food Policy*, 121:102539.

Tallón-Ballesteros, A. and Chen, C. (2020). Explainable ai: Using shapley value to explain complex anomaly detection ml-based systems. *Machine learning and artificial intelligence*, 332:152.

Tang, B., Liu, Y., and Matteson, D. S. (2022). Predicting poverty with vegetation index. *Applied Economic Perspectives and Policy*, 44(2):930–945.

Tennant, E., Ru, Y., Sheng, P., Matteson, D. S., and Barrett, C. B. (2025). Microlevel structural poverty estimates for southern and eastern Africa. *Proceedings of the National Academy of Sciences*, 122(6):e2410350122.

Upton, J., Tennant, E., Fiorella, K. J., and Barrett, C. B. (2023). Covid-19, household resilience, and rural food systems: evidence from southern and eastern africa. In *Resilience and food security in a food systems context*, pages 281–320. Springer International Publishing Cham.

van der Weide, R., Blankespoor, B., Elbers, C., and Lanjouw, P. (2024). How accurate is a poverty map based on remote sensing data? An application to Malawi. *Journal of Development Economics*, 171:103352.

Victora, C. G., Adair, L., Fall, C., Hallal, P. C., Martorell, R., Richter, L., and Sachdev, H. S. (2008). Maternal and child undernutrition: Consequences for adult health and human capital. *The Lancet*, 371(9609):340–357.

Wiesmann, D., Bassett, L., Benson, T., and Hoddinott, J. F. (2009). Validation of the world food programme's food consumption score and alternative indicators of household food security. Technical Report 00870, International Food Policy Research Institute (IFPRI), Washington, D.C.

World Bank (2021). The Human Capital Index 2020 Update: Human Capital in the Time of COVID-19. Technical report, Washington, DC: World Bank.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1):2583.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American statistical Association*, 105(489):312–323.

Zheng, Z., Wu, T., Lee, R., Newhouse, D., Kilic, T., Burke, M., Ermon, S., and Lobell, D. B. (2025). Dynamic, high-resolution wealth measurement in data-scarce environments. *World Bank Policy Research Working Paper*.

Zhou, Y., Lentz, E., Michelson, H., Kim, C., and Baylis, K. (2022). Machine learning for food security: Principles for transparency and usability. *Applied Economic Perspectives and Policy*, 44(2):893–910.