Discussion Paper Series

IZA DP No. 18893

August 2026

Shelter from the Storm: A Simulation Framework for Vulnerability under Climate Shocks

Gustavo Canavire-Bacarreza
World Bank, Universidad Privada Boliviana and IZA@LISER

Alejandro Puerta-Cuartas
Universidad Carlos III

Carlos Rodriguez-Castelán
World Bank and IZA@LISER

Carolina Vélez-Ospina
World Bank

IZA@LISER NETWORK
ISSN: 2365-9793

IZA DP No. 18893 August 2026

Shelter from the Storm: A Simulation Framework for Vulnerability under Climate Shocks*

Abstract

This paper proposes a nonparametric simulation framework to estimate poverty vulnerability under climate shocks. We formalize vulnerability estimation as an out-of-sample prediction problem and show that flexible, regularized machine learning methods for estimating the conditional mean of welfare offer a powerful alternative to conventional linear models. The framework simulates future welfare distributions using historical realizations of climate shocks and household characteristics, enabling the estimation of vulnerability measures and related functions without imposing restrictive parametric assumptions. To interpret the model and quantify heterogeneous impacts, we employ SHapley Additive exPlanations, which decompose predicted vulnerability into contributions from climate shocks and household characteristics. An application to Ecuador reveals a strong geographic concentration of vulnerability and shows that climate shocks act as localized triggers that push marginal households, particularly low-educated informal rural workers into poverty.

JEL classification

I32, I38, C14, C15

Keywords

poverty vulnerability, climate shocks, machine learning

Corresponding author

Gustavo Canavire-Bacarreza
gcanavire@worldbank.org

* The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the World Bank, its Board of Directors, or the countries it represents. The authors declare the use of Generative AI for language editing and paraphrasing to improve clarity. We thank participants at the ECINEQ-LAC Meeting in Mexico City and IAAE Annual Conference in Carcavelos for their helpful comments and suggestions. All remaining errors are of the authors.

1 Introduction

Recent decades have witnessed a marked increase in the frequency and severity of climate-related shocks. Floods, droughts, and temperature extremes are occurring more often and with greater intensity, disrupting economic activity and threatening household welfare across many parts of the world (Cred, 2020; World Meteorological Organization, 2021). These shocks affect agricultural productivity, labor markets, and local prices, and their impacts are often disproportionately borne by households with limited economic opportunities and fewer resources to cope with adverse events (Dercon, 2002; Hallegatte et al., 2017). As climate risks intensify, understanding which households are most exposed to falling into poverty has become an increasingly important challenge for policymakers seeking to design preventive and adaptive policy responses (Hallegatte et al., 2016).

Traditional poverty measures capture realized deprivation but provide limited information about the risk that households currently above the poverty line may fall into poverty in the future—that is, their vulnerability to poverty (Ligon and Schechter, 2003). Estimating vulnerability is challenging because the effect of shocks on household welfare depends not only on exposure to climate events but also on household characteristics and local economic conditions. These interactions are likely to be nonlinear and heterogeneous across regions and socioeconomic groups, making vulnerability difficult to measure using conventional empirical approaches.

A large literature has sought to address this challenge by estimating the probability that households will fall below the poverty line in the future. Early contributions, including Chaudhuri (2000), proposed cross-sectional methods that approximate the expected distribution of future consumption or income using parametric models. Subsequent work has expanded this framework in several directions, including the simulation-based approach proposed by Hill and Porter (2017), which estimates vulnerability by simulating future welfare outcomes under alternative realizations of shocks. This approach has gained prominence because it allows researchers to analyze vulnerability even in contexts where long household panels are unavailable (Dercon, 2002).

Despite these advances, existing empirical implementations often rely on restrictive parametric assumptions. In particular, the conditional mean of welfare is typically estimated using linear models, which may fail to capture the nonlinearities and heterogeneous responses that characterize the relationship between household characteristics, environmental shocks, and economic outcomes. These limitations are particularly relevant in the context of climate shocks, where the impacts of environmental variation often interact with local economic structures, geographic conditions, and household characteristics in complex ways (Hallegatte et al., 2017).

This paper proposes a nonparametric simulation framework for analyzing poverty vulnerability, understood as the probability of being poor, with particular attention to the role of climate shocks and the design of targeted, place-based public policies. For this purpose, we formalize and generalize the simulation approach to climate vulnerability of Hill and Porter (2017) (the Hill–Porter approach, HPA henceforth). Specifically, we show that the HPA can be formally interpreted as an out-of-sample prediction problem. While HPA focus on estimating vulnerability measures themselves, we demonstrate that the framework naturally extends to the estimation of functions of these measures. These functions include, for example, the characteristics of the most vulnerable populations and the heterogeneous impacts of specific climate shocks on poverty vulnerability.

The HPA estimates poverty vulnerability by simulating the probable distribution of a future welfare aggregate—such as consumption or income—using historical data on shocks. The procedure consists of three steps. First, the conditional mean function $m(\cdot)$ relating the welfare aggregate to household characteristics and shocks is estimated using ordinary least squares (OLS). Second, the future welfare aggregate is simulated using historical realizations of shocks together with the estimated conditional mean function $\hat{m}(\cdot)$. Third, vulnerability measures are computed based on the simulated distribution of the future welfare aggregate. Although the HPA considers several types of shocks—including death, job loss, price shocks, and climate shocks—this paper focuses on the last category.

To uncover the most salient features of vulnerable households and estimate the heterogeneous impact of climate shocks on future welfare, we employ SHapley Additive exPlanations (SHAP) as proposed by Lundberg (2017). SHAP values provide a measure of feature importance that allows us to identify the characteristics most strongly associated with vulnerability. In addition, SHAP values quantify the contribution of each explanatory variable to the predicted welfare outcome at the observation level. This feature makes it possible to evaluate the heterogeneous effects of climate shocks on the simulated future welfare aggregate.

The main contribution of this paper is to present an easy-to-implement procedure for analyzing poverty vulnerability, assessing the impact of climate shocks, and designing targeted place-based policies. Our proposal consists of four stages: (1) estimating the conditional mean function relating consumption (or income) to household characteristics and climate shocks using ML methods; (2) simulating the future welfare aggregate using household characteristics and historical climate shocks together with the estimated conditional mean function; (3) predicting vulnerability status based on the simulated future welfare aggregate; and (4) estimating functions of vulnerability. When the function of interest is known, we use its sample counterpart to estimate it.[1] Conversely, when the function is unknown, we propose computing SHAP values from a classification model that regresses non-parametrically the estimated vulnerability status on household characteristics and climate shocks.

Our method can be applied to both short panels and cross-sectional data. Similar to multilevel approaches to vulnerability estimation (for example, Skoufias and Quisumbing (2003) and Günther and Harttgen (2009)) and simulation-based approaches (for example, Hill and Porter (2017)), our framework does not require long panels of income or consumption data. Instead, it requires only household survey data containing socioeconomic characteristics combined with historical climate data at the community or regional level. In our empirical application we rely on publicly available climate data, illustrating that practitioners need only merge survey data with such data sources in order to implement the proposed procedure.

We illustrate our framework with an empirical application to Ecuador, a country characterized by both high climate exposure and substantial socioeconomic vulnerability. Ecuador is one of the most biodiverse countries in the world (Kleemann et al., 2022; Canavire-Bacarreza et al., 2026; Canavire-Bacarreza and Serio, 2026), with 96 percent of its population living in coastal and mountainous regions exposed to multiple natural hazards (World Bank, 2014). According to the Survey of Living Conditions (Encuesta de Condiciones de Vida) conducted by the Instituto Nacional de Estadística y Censos (INEC), approximately one out of four households faces the risk of poverty or social exclusion (INEC, 2023).

To conduct the empirical analysis, we construct a panel covering the period 2007–2021 that combines climate and household data. We consider four types of climate shocks, maximum temperature, minimum temperature, floods, and droughts. Climate data are obtained from the Climatic Research Unit of the University of East Anglia (Harris et al., 2020),[2] which provides gridded time-series information at a spatial resolution of approximately 1 km^2. These data are merged with the national household survey conducted by INEC, yielding a pooled dataset containing nearly 188,000 household heads.

Our empirical application to Ecuador reveals that poverty vulnerability has become

[1]For instance, to characterize vulnerable populations we can compute averages for binary variables and the empirical cumulative distribution function (ECDF) for continuous variables. Provided information on rural residence and age, we can also estimate the proportion of vulnerable workers in the primary sector and the age distribution of the vulnerable.

[2]The data can be accessed at `https://www.worldclim.org/data/`.

increasingly concentrated in the Amazonian region over the study period, with Napo ranking first in 2021, while several Andean provinces such as Bolívar and Chimborazo exhibited persistently high vulnerability in earlier years. Coastal provinces consistently rank among the least vulnerable throughout 2007–2021. The socioeconomic profile of the vulnerable is multidimensional, informality, self-employment, low education, primary-sector employment, rural residence, and age all contribute similarly to predicted vulnerability, with no single characteristic dominating. Household-level conditions together account for roughly 70 percent of the explained variation in vulnerability. These findings motivate a formalization policy targeting self-employed primary-sector workers with low educational attainment in rural areas, particularly those aged 25 to 37. Although climate shocks play a secondary role relative to structural socioeconomic factors in the aggregate decomposition, their impacts are spatially concentrated and can be severe at the local level. Floods primarily affect provinces along the coast and Andean corridor, droughts concentrate in the Andean highlands, and extreme heat is most severe in the Amazon. Moreover, within a given province the impact of a shock varies substantially across households. These results suggest that climate-related poverty risks are best addressed through interventions that are both geographically targeted and tailored to the socioeconomic profile of the most exposed populations, rather than through uniform national policies.

The remainder of the paper proceeds as follows. Section 2 formalizes the simulation approach proposed by Hill and Porter (2017) as an out-of-sample prediction problem and motivates the adoption of a machine learning approach for simulating the future welfare aggregate. Section 3 generalizes the simulation approach to climate vulnerability by extending it to the prevention or mitigation of climate risk and presents the proposed estimation framework. Section 4 presents the empirical application to Ecuador, and Section 5 concludes.

2 A Formalization of the Simulation Approach to Climate Vulnerability

Our point of departure is the estimation of a measure of vulnerability $\mathcal{V}$, which depends on the future welfare aggregate C_{t+1}, the poverty line z, and an exogenous parameter α. The welfare aggregate may correspond to consumption or income depending on the empirical application; we retain the notation C to align with the HPA. Accordingly, our object of interest can be defined as

$$\mathcal{V} \coloneqq g\left(C_{t+1}, z, \alpha\right), \tag{1}$$

where the function $g(\cdot)$ is known up to C_{t+1}, z, and α. Thus, if the distribution of the future welfare aggregate were known, the vulnerability measure could be computed directly given values for z and α.

Equation (1) encompasses a wide class of vulnerability measures. Consider first the expected poverty measure of Chaudhuri et al. (2002),

$$\mathcal{V} = g\left(C_{t+1}, z, \alpha\right) = \mathbb{E}\left[\left(\max\left\{0, \frac{z - C_{t+1}}{z}\right\}\right)^{\alpha}\right],$$

where the parameter α determines whether the object of interest is the headcount ($\alpha = 0$), the poverty gap ($\alpha = 1$), or the squared gap ($\alpha = 2$). A second example is the measure of Pritchett et al. (2000), in which the probability that a household is vulnerable increases with the probability that its future welfare aggregate falls below the poverty line, so that equation (1) reduces to

$$\mathcal{V} = g\left(C_{t+1}, z, \alpha\right) = P\left(C_{t+1} \leq z\right) > \alpha,$$

where α is commonly set to 0.5. Finally, adopting the constant relative risk-sensitive measure of Calvo and Dercon (2013) yields a vulnerability measure of the form

$$\mathcal{V} = g\left(C_{t+1}, z, \alpha\right) = \frac{1}{\alpha}\left(1 - \mathbb{E}\left[\left(\frac{C_{t+1}}{z} \times 1\left(C_{t+1} < z\right) + z \times 1\left(C_{t+1} \geq z\right)\right)^{\alpha}\right]\right).$$

Because the future welfare aggregate is unknown, it must be estimated in order to measure vulnerability. The HPA addresses this by simulating the probable distribution of the future welfare aggregate conditional on household characteristics and historical data on shocks. In this paper we focus on climate shocks. To formalize the approach for this type of shock, we decompose the welfare aggregate into observed and unobserved components as follows:

$$C = m\left(\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\eta}\right) + \epsilon, \quad \mathbb{E}\left[\epsilon | \boldsymbol{X}, \boldsymbol{S}\right] = 0, \tag{2}$$

where $\boldsymbol{X}$ and $\boldsymbol{S}$ denote household characteristics and climate shocks, respectively; $m(\cdot) \coloneqq \mathbb{E}\left[C | \boldsymbol{X}, \boldsymbol{S}\right]$ is the conditional mean function, known up to the nuisance parameter $\boldsymbol{\eta}$;[3] and ϵ is an idiosyncratic shock corresponding to the unexplained component.

The HPA is motivated by the observation that if $m(\cdot)$, $\boldsymbol{\eta}$, and the distributions of the climate and idiosyncratic shocks were known, the future welfare aggregate could be

[3] For instance, if we assume that $m(\cdot)$ is linear, the nuisance parameter $\boldsymbol{\eta}$ corresponds to the regression coefficients. We provide a characterization of the conditional mean function and the nuisance parameter for the case of tree-based algorithms and kernel machines in Appendix A.

simulated as

$$\begin{aligned} C_{t+1}^{m} &= m\left(\boldsymbol{X},\boldsymbol{S}^{m};\boldsymbol{\eta}\right)+\epsilon^{m}, \\ S^{m} &\sim F_{S}, \\ \epsilon^{m} &\sim F_{\epsilon}, \end{aligned} \tag{3}$$

where F_S and F_ϵ are the distributions of the climate and idiosyncratic shocks, respectively, and S^m and ϵ^m are draws from them. By drawing C_{t+1}^m a sufficiently large number of times (say M), one can approximate the distribution function of the future welfare aggregate and, in turn, compute the vulnerability measure in equation (1).

The HPA can be formalized as a three-step procedure. The first step estimates the equation relating the welfare aggregate (C) to household characteristics ($\boldsymbol{X}$) and climate shocks ($\boldsymbol{S}$),

$$C = m\left(\boldsymbol{X},\boldsymbol{S};\boldsymbol{\eta}\right)+\epsilon, \quad \mathbb{E}\left[\epsilon|\boldsymbol{X},\boldsymbol{S}\right]=0,$$

by OLS, assuming

$$m\left(\boldsymbol{X},\boldsymbol{S};\boldsymbol{\eta}\right)=\boldsymbol{X}'\boldsymbol{\beta}+\boldsymbol{S}'\boldsymbol{\delta}+\boldsymbol{S}'\tilde{\boldsymbol{X}}'\boldsymbol{\gamma}, \tag{4}$$

where $\tilde{\boldsymbol{X}}$ is a subset of $\boldsymbol{X}$, so that the effect of climate shocks on consumption is allowed to depend on household characteristics. The estimated conditional mean function is then

$$\hat{m} \coloneqq m\left(\boldsymbol{X},\boldsymbol{S};\hat{\boldsymbol{\eta}}\right)=\tilde{\boldsymbol{Z}}'\left(\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}'\right)^{-1}\tilde{\boldsymbol{Z}}'C \coloneqq \tilde{\boldsymbol{Z}}'\hat{\boldsymbol{\eta}}, \quad \tilde{\boldsymbol{Z}}' \coloneqq \left(\boldsymbol{X}',\boldsymbol{S}',\boldsymbol{S}'\tilde{\boldsymbol{X}}'\right)',$$

and the ECDF of the residuals $\hat{\epsilon} \coloneqq C-\tilde{\boldsymbol{Z}}'\hat{\boldsymbol{\eta}}$ is given by

$$\hat{F}_{\epsilon}(e)=\frac{1}{n}\sum_{i=1}^{n}1\left(C_i-\tilde{\boldsymbol{Z}}_i'\hat{\boldsymbol{\eta}}<e\right) \coloneqq \mathbb{E}_n\left[1\left(C_i-m\left(\boldsymbol{X}_i,\boldsymbol{S}_i;\hat{\boldsymbol{\eta}}\right)<e\right)\right],$$

where $\mathbb{E}_n\left[\cdot\right]$ is the empirical expectation operator.

The second step simulates the future welfare aggregate M times by drawing climate and idiosyncratic shocks from their respective ECDFs,

$$\begin{aligned} \hat{C}_{t+1}^{m} &= m\left(\boldsymbol{X},\boldsymbol{S}^{m};\hat{\boldsymbol{\eta}}\right)+\epsilon^{m}=\tilde{\boldsymbol{Z}}^{m\prime}\hat{\boldsymbol{\eta}}+\epsilon^{m}, \quad m=1,...,M, \\ S^{m} &\sim \hat{F}_{S}, \quad \hat{F}_S(s)=\mathbb{E}_n\left[1\left(\boldsymbol{S}_i<s\right)\right], \\ \epsilon^{m} &\sim \hat{F}_{\epsilon}, \quad \hat{F}_{\epsilon}(e)=\mathbb{E}_n\left[1\left(C_i-m\left(\boldsymbol{X}_i,\boldsymbol{S}_i;\hat{\boldsymbol{\eta}}\right)<e\right)\right], \end{aligned} \tag{5}$$

where $\tilde{\boldsymbol{Z}}^{m'} := \left(\boldsymbol{X}', \boldsymbol{S}^{m'}\right)'$. From these draws one recovers the distribution of the future welfare aggregate,

$$\hat{C}_{t+1} \sim \hat{F}_C, \quad \hat{F}_C(c) = \frac{1}{M} \sum_{m=1}^{M} 1\left(\hat{C}_{t+1}^{m} < c\right),$$

so that

$$\hat{C}_{t+1} = f\left(C, \boldsymbol{X}, \phi, m\right), \quad \phi := \left(F_S, F_\epsilon\right). \tag{6}$$

The simulated future welfare aggregate thus depends on observed consumption, household characteristics, the distributions of the climate and idiosyncratic shocks, and the conditional mean function.

The third step estimates the vulnerability measure in equation (1) as

$$\hat{\mathcal{V}} = g\left(\hat{C}_{t+1}, \alpha, z\right). \tag{7}$$

This procedure requires estimating both a conditional mean function and a set of unknown distribution functions. To see this, substitute equation (6) into equation (7) and reparametrize the object of interest as

$$\hat{\mathcal{V}} = g\left(\hat{C}_{t+1}, z, \alpha\right) := \tilde{f}\left(C, \boldsymbol{X}, \phi, m, z, \alpha\right).$$

The estimated vulnerability measure therefore depends on consumption and household characteristics $(C, \boldsymbol{X})$, the unknown distribution functions (ϕ), the conditional mean function (m), and the exogenous values (z, α). Because the data and the poverty line are given and α is fixed ad hoc for the analysis, only the conditional mean and the unknown distribution functions remain to be estimated.

The HPA estimates the distributions of the future welfare aggregate and of the climate and idiosyncratic shocks with the ECDF. This choice is justified by the Glivenko–Cantelli theorem, which establishes that

$$||F_n - F||_\infty := \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0 \;\; a.s., \tag{8}$$

where F is the common cumulative distribution of the independent and identically distributed random variables $X_1, ..., X_M$ in $\mathbb{R}$, and

$$F_n(x) \coloneqq \frac{1}{M} \sum_{m=1}^{M} 1\left(X_m < x\right)$$

is the ECDF of $\{X_i\}_{i=1}^{M}$ evaluated at x. Equation (8) shows that the ECDF converges uniformly and almost surely to the true distribution function, so that it is a consistent estimator of the cumulative distribution function (CDF). This justifies its use within the HPA to estimate the unknown distribution functions.

The defining feature of this procedure is that estimating the conditional mean function serves the purpose of out-of-sample prediction. As equation (5) makes explicit, simulating the future welfare aggregate requires evaluating $\hat{m}(\cdot)$ at climate shocks drawn from their empirical distribution—that is, at test observations that were not used in estimation. Two implications follow. First, because the effect of climate shocks on welfare is likely to be heterogeneous across households and regions, $m(\cdot)$ is potentially nonlinear, and its estimator must be flexible enough to capture such nonlinearities. Second, because the object of estimation is a prediction, the procedure must incorporate regularization to guard against overfitting. Estimating $m(\cdot)$ by OLS is unsatisfactory on both counts: it imposes linearity and provides no regularization, entailing a substantial loss in the predictive accuracy of the simulated welfare aggregate.

The linearity of OLS can be partially relaxed by projecting the covariates into a higher-dimensional space through basis functions—for instance, mapping each covariate into the space of powers $t(x) = (1, x, x^2, x^3, ...)'$ or introducing interactions among covariates in the spirit of equation (4). As long as these functions do not depend on the parameters, the model remains linear in the parameters (Williams and Rasmussen, 2006).

The HPA relaxes linearity through this second route, interacting a subset of household characteristics with climate shocks on the basis of economic reasoning. While selecting interactions a priori helps limit overfitting by omitting variables, the choice is not data-driven and may entail a substantial loss in accuracy. More fundamentally, this strategy allows the effect of shocks to vary across households but not across regions. Yet the extent to which a shock affects welfare depends on both socioeconomic attributes and the geographic characteristics of the region of residence: certain mechanisms are inherently tied to specific household characteristics (an indirect effect), whereas the forces leading to poverty are partly external (a direct effect). Regional heterogeneity could in principle be accommodated by interacting climate shocks with regional dummies, but doing so would sharply increase the number of parameters and, with it, the need for regularization.

Adopting an ML approach directly accommodates the predictive nature of simulating future consumption (or income). ML methods model the conditional mean function nonparametrically, relaxing the linearity assumption of OLS, while simultaneously incorporating regularization that guards against overfitting. In this sense, our formalization of the HPA motivates the use of ML to simulate the future welfare aggregate and, more generally, welfare aggregates of interest.

Despite these advantages, the ML approach is not without cost: it induces bias in the predicted vulnerability status, arising from the regularization and model-selection bias inherent in ML estimators of the conditional mean (Chernozhukov et al., 2022). Errors in the out-of-sample prediction of the welfare aggregate propagate to the estimated vulnerability status, so that households may be misclassified as vulnerable when their future welfare aggregate is poorly approximated. This limitation, however, is not unique to ML. OLS is subject to analogous problems for different reasons: the linearity assumption yields a poor approximation of the welfare aggregate, and the inclusion of numerous interactions without regularization can induce overfitting (Hawkins, 2004).

3 A Nonparametric Simulation Framework for Poverty Vulnerability and Climate Risk

3.1 Estimating Functions of the Future Welfare Aggregate

Equation (1) shows that a measure of vulnerability can be estimated by simulating the future welfare aggregate. We now generalize this object of interest to a broader setting in which the target is a function of vulnerability. Specifically, we focus on the estimation of a parameter θ, defined by a function $h(\cdot)$ of household characteristics, climate shocks, the welfare aggregate, the poverty line z, and an exogenous parameter α:

$$\theta = h\left(\boldsymbol{Z}, C_{t+1}, z, \alpha\right), \tag{9}$$

where $\left(\boldsymbol{Z} \coloneqq (\boldsymbol{X}', \boldsymbol{S}')'\right)$. We first consider a particular case of equation (9) in which the object of interest depends on household characteristics, climate shocks, and the household vulnerability status V_i, determined according to the vulnerability measure of Pritchett et al. (2000):

$$V_i = 1\left(P\left(C_{i,t+1} \leq z\right) > \alpha\right), \quad i = 1, ..., n. \tag{10}$$

Our object of interest then takes the form

$$\theta = h\left(\boldsymbol{Z}, C_{t+1}, z, \alpha\right) = \tilde{h}\left(\boldsymbol{Z}, V\right), \tag{11}$$

where the second equality uses equations (9) and (10), and V is a vector recording the vulnerability status of every household.

Equation (11) generalizes the objects of interest in Hill and Porter (2017). In their empirical application, those authors estimate five vulnerability measures by gender and by rural versus urban residence, which can be formalized as

$$\theta = \tilde{h}\left(\boldsymbol{Z}, V\right) = \mathcal{V}|\boldsymbol{Z},$$

where $\boldsymbol{Z}$ is the conditioning variable (gender and rural versus urban residence).[4]

The main implication of equation (11) is that the HPA can be used to estimate functions of vulnerability and, as we now show, to inform the design of targeted, place-based policies for preventing and mitigating climate risk. We illustrate this with a motivating example. Suppose the object of interest in equation (11) is

$$\theta = \tilde{h}\left(\boldsymbol{Z}, V\right) = F_{\boldsymbol{X}|V=1}\left(\boldsymbol{x}\right) = P\left(X_1 \leq x_1, ..., X_r \leq x_{ls}|V=1\right), \tag{12}$$

where $F_{\boldsymbol{X}|V=1}\left(\boldsymbol{x}\right)$ is the joint CDF of the characteristics of the vulnerable and l_s is the cardinality of $\boldsymbol{X}$. If household characteristics comprise only age (age) and a dummy ($prim$) equal to one for primary-sector employment and zero otherwise, equation (12) reduces to

$$F_{\boldsymbol{X}|V=1}\left(\boldsymbol{x}\right) = P\left(age_1 \leq x_1, prim \leq x_2|V=1\right),$$

with marginal distributions

$$F_{age}(a) = \mathbb{E}\left[1\left(age \leq a\right)|V=1\right],$$

and

$$F_{prim}(a) = \mathbb{E}\left[1\left(prim \leq a\right)|V=1\right].$$

Were vulnerability found to concentrate among older primary-sector workers, the government could target climate risk mitigation efforts toward that population.

[4]Note that the vulnerability measure $\mathcal{V}$ summarizes in a scalar the vector V, which records the vulnerability status of each household.

To operationalize the identification of the most salient characteristics of the vulnerable, we estimate equation (10) using the first two steps of the HPA. Following equation (5), this yields

$$\hat{V}_i = 1\left(P\left(\hat{C}_{i,t+1} \leq z\right) > \alpha\right) \coloneqq h\left(C_i, \boldsymbol{X}_i, \hat{\phi}, \hat{m}, z, \alpha\right), \quad \hat{\phi} \coloneqq \left(\hat{F}_S, \hat{F}_\epsilon\right), \tag{13}$$

so that the household vulnerability status can be expressed as a function of household characteristics (among other arguments). This follows because simulating the future welfare aggregate under the HPA holds household characteristics fixed (see equation (5)). Accordingly, to identify the most salient characteristics of the vulnerable we estimate

$$\theta = \tilde{h}\left(\boldsymbol{Z}, V\right) = \tilde{h}\left(\boldsymbol{X}, V\right) = \nabla_{\boldsymbol{X}} V_i = \begin{bmatrix} \frac{\partial P(V_i=1)}{\partial X_1} \\ \vdots \\ \frac{\partial P(V_i=1)}{\partial X_{l_x}} \end{bmatrix}, \quad i = 1, ...n, \tag{14}$$

where $\nabla_{\boldsymbol{X}} V_i$ is the gradient of the household probability of vulnerability with respect to household characteristics.

Equation (14) captures the relationship between vulnerability status and household characteristics, and its usefulness is twofold. First, it provides a measure of variable importance: the characteristics with the largest impact on the probability of being vulnerable are the most salient attributes of the vulnerable. Second, it characterizes the most vulnerable households. If $\frac{\partial P(V_i=1)}{\partial Age}$ were found to be U-shaped, for instance, the youngest and oldest would be the most vulnerable; if the derivative were constant and negative, the youngest would be.

The HPA simulation framework treats its inputs asymmetrically, fixing household characteristics $\boldsymbol{X}_i$ at their observed values while drawing climate shocks from their empirical distribution $\hat{F}_S$ rather than using the realized shocks $\boldsymbol{S}_i$. This asymmetry enables the characteristic-based analysis of equation (14), but it introduces a fundamental limitation for shock analysis: the vulnerability measures $\hat{V}_i$ come to depend on the shock distribution $\hat{F}_S$ rather than on specific realizations, effectively averaging over all possible shock combinations. This averaging masks the distinct effects of individual shocks and prevents direct measurement of how estimated vulnerability responds to any particular one.

To overcome this limitation, we exploit the realized climate shocks that households actually experienced. To analyze the effect of a particular shock S_k, we fix its observed value and simulate the remaining shocks from their empirical distribution. Formally, for each climate shock variable S_k (with $k = 1, \ldots, l_s$), we fix each household i's observed exposure S_{ik} to shock k, simulate M realizations of the remaining shocks $\boldsymbol{S}_{-k}$ from their joint empirical distribution $\hat{F}_{S_{-k}}$, and, for each draw $m = 1, \ldots, M$, compute the future welfare aggregate as

$$\hat{C}^{m}_{i,t+1} = m(\boldsymbol{X}_i, S_{ik}, \boldsymbol{S}^{m}_{-k}; \hat{\boldsymbol{\eta}}) + \epsilon^m,$$

where $\boldsymbol{X}_i$ denotes household characteristics, $\hat{\boldsymbol{\eta}}$ the estimated model parameters, and ϵ^m an error term. This construction makes the estimated vulnerability measure for each household depend on its characteristics $\boldsymbol{X}_i$, the realized shock S_{ik}, and the distribution of the remaining shocks. It thus preserves the structural relationship between shocks and consumption in equation (5) while isolating the contribution of each shock to poverty vulnerability, in the spirit of equation (14). To capture the heterogeneous impacts of different climate shocks, we then estimate the gradient

$$\theta = \tilde{h}(\boldsymbol{Z}, V) = \tilde{h}(\boldsymbol{S}, V) = \nabla_{\boldsymbol{S}} V_i = \begin{bmatrix} \frac{\partial P^1(V_i=1)}{\partial S_1} \\ \vdots \\ \frac{\partial P^{l_s}(V_i=1)}{\partial S_{l_s}} \end{bmatrix}, \quad i = 1, \ldots, n, \tag{15}$$

where $P^k(V_i = 1)$ is the probability of vulnerability for household i conditional on its characteristics $\boldsymbol{X}_i$ and the realized shock S_{ik}, with the remaining climate shocks and the idiosyncratic shock simulated from their empirical distributions.

Like equation (14), equation (15) identifies who is most affected when a specific climate shock occurs. Because the derivatives are computed at the observation level and socioeconomic characteristics are observed, we can pinpoint the populations and regions most affected by each shock, enabling governments to concentrate preventive action where it is most needed. Local governments in flood-prone regions, for example, can prioritize clearing debris and waste from drainage systems to reduce blockages, while regions most exposed to extreme heat can invest in cool roofs and expanded urban greenery.[5]

In sum, this subsection has generalized the HPA to the estimation of functions of the future welfare aggregate and has shown how estimating equations (14) and (15) supports the design of targeted, place-based policies to reduce poverty vulnerability. We now describe our proposal for estimating these two equations.

[5]For evidence on the cooling effects of cool roofs and urban greenery, see Macintyre and Heaviside (2019) and Đekić et al. (2018), respectively.

3.2 Estimation Procedure

In Section 2 we highlighted the benefits of modeling the welfare aggregate as a function of household characteristics and climate shocks with ML. Before presenting the estimation procedure, we show how the ML approach can be complemented with tools from the explainable artificial intelligence literature to reduce poverty risk. In particular, equations (14) and (15) can be estimated by computing the SHAP values proposed by Lundberg (2017).

The SHAP values build upon the Shapley values (Shapley, 1953), which prescribe how the surplus of a cooperative game should be allocated among its players. Specifically, in a coalition of N agents, the Shapley value of agent j is the fair share of the coalition's value (V) that j should receive,

$$\phi_j\left(V\right) = \frac{1}{N}\sum_{S}\frac{\left[V\left(S\cup\{j\}\right) - V\left(S\right)\right]}{\binom{N-1}{k_s}},$$

where the summation runs over all subsets S of the team $T = \{1, ..., N\}$ that can be formed after excluding j, k_s is the number of agents in coalition S, $V(S)$ is the value achieved by subteam S, and $V\left(S\cup\{j\}\right)$ is the value realized once j joins S. Thus $\phi_j\left(V\right)$ measures the average contribution of j, which constitutes its fair share.

In a predictive setting, the Shapley values evaluate the effect of each variable on the prediction of an outcome for every observation. The agents correspond to the explanatory variables and the coalitions to models fitted on a subset of them: intuitively, each variable's Shapley value is obtained by estimating every possible model with and without that variable and comparing their predictions of the dependent variable. Because the number of such combinations is large, the SHAP values rely on an approximation of the Shapley values (for a detailed formulation, see Lundberg (2017)).

We therefore propose computing SHAP values to estimate equations (14) and (15). In our framework, the SHAP values measure the impact of each variable on vulnerability and welfare outcomes, providing estimates of both objects of interest. By computing them, we can characterize the households most vulnerable to climate shocks and estimate the heterogeneous effect of specific climate shocks on their future welfare.

Our proposal for estimating functions of vulnerability proceeds in four steps. The first step estimates the equation relating the welfare aggregate (C) to household characteristics $(\boldsymbol{X})$ and climate shocks $(\boldsymbol{S})$,

$$C = m\left(\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\eta}\right) + \epsilon, \quad \mathbb{E}\left[\epsilon|\boldsymbol{X}, \boldsymbol{S}\right] = 0,$$

by ML, yielding the nonparametric estimate of the conditional mean $m(\boldsymbol{X}, \boldsymbol{S}; \hat{\boldsymbol{\eta}})$ and the ECDF of the residuals $\hat{\epsilon} := C - m(\boldsymbol{X}, \boldsymbol{S}; \hat{\boldsymbol{\eta}})$,

$$\hat{F}_\epsilon(e) = \frac{1}{n}\sum_{i=1}^{n} 1\left(C_i - m\left(\boldsymbol{X}_i, \boldsymbol{S}_i; \hat{\boldsymbol{\eta}}\right) < e\right) := \mathbb{E}_n\left[1\left(C_i - m\left(\boldsymbol{X}_i, \boldsymbol{S}_i; \hat{\boldsymbol{\eta}}\right) < e\right)\right].$$

The second step simulates the future welfare aggregate M times by drawing climate and idiosyncratic shocks from their ECDFs,

$$\begin{aligned}
\hat{C}^m_{t+1} &= m\left(\boldsymbol{X}, \boldsymbol{S}^m; \hat{\boldsymbol{\eta}}\right) + \epsilon^m, \quad m = 1, ..., M, \\
S^m &\sim \hat{F}_S, \quad \hat{F}_S(s) = \mathbb{E}_n\left[1\left(\boldsymbol{S}_i < s\right)\right], \\
\epsilon^m &\sim \hat{F}_\epsilon, \quad \hat{F}_\epsilon(e) = \mathbb{E}_n\left[1\left(C_i - m\left(\boldsymbol{X}_i, \boldsymbol{S}_i; \hat{\boldsymbol{\eta}}\right) < e\right)\right].
\end{aligned}$$

When the interest lies in a specific shock, the fixed shock S_k is held at its realized value while the remaining shocks are drawn from their joint empirical distribution,

$$\hat{C}^m_{t+1} = m\left(\boldsymbol{X}, S_k, \boldsymbol{S}^m_{-k}; \hat{\boldsymbol{\eta}}\right) + \epsilon^m, \quad \boldsymbol{S}^m_{-k} \sim \hat{F}_{S_{-k}}.$$

From these draws one recovers the distribution of the future welfare aggregate,

$$\hat{C}_{t+1} \sim \hat{F}_C, \quad \hat{F}_C(c) = \frac{1}{M}\sum_{m=1}^{M} 1\left(\hat{C}^m_{t+1} < c\right).$$

The third step computes the household vulnerability status,

$$\hat{V}_i = 1\left(P\left(\hat{C}_{i,t+1} \leq z\right) > \alpha\right), \quad i = 1, ..., n,$$

and the fourth step estimates equation (11) as

$$\hat{\theta} = \hat{\tilde{h}}\left(\boldsymbol{Z}, \hat{V}\right), \quad \hat{V} = \left(\hat{V}_1, ..., \hat{V}_n\right).$$

When $\tilde{h}$ is known—for example, a conditional mean or a vulnerability measure for a subgroup—the estimator $\hat{\tilde{h}}$ is simply the sample counterpart of h. When h is unknown, we estimate it using the SHAP values from a classification model that regresses the estimated vulnerability status on household characteristics and climate shocks.

Importantly, in the second step the remaining shocks $\boldsymbol{S}^m_{-k}$ are drawn from their joint empirical distribution $\hat{F}_{S_{-k}}$ rather than independently from their marginals. This preserves the correlation structure among the non-fixed climate variables, mitigating unrealistic scenarios within a given location and time period.

The main differences between our proposal and the HPA lie in the first and fourth steps. Rather than estimating the conditional mean by OLS, the first step performs nonparametric estimation by ML. The second and third steps coincide with those of the HPA, except that the third step retains the household vulnerability status instead of collapsing it into a vulnerability measure. The fourth step, finally, encompasses a wide class of objects of interest, including vulnerability measures for the whole population and for subpopulations (such as male versus female or rural versus urban residence, as in Hill and Porter (2017)) as well as equation (14).

A critical note concerns the first step: because ML methods are used, practitioners must conduct hyperparameter tuning to avoid overfitting. Care must be taken to ensure that this process does not assign zero weight to the climate shock variables, since doing so would restrict the second step to idiosyncratic shocks and household characteristics alone. We perform the tuning by combining k-fold cross-validation with Bayesian optimization through a probabilistic grid search (Shcherbatyi et al., 2024), using the mean squared error as the scoring criterion. The chief advantage of Bayesian optimization is that it models the objective function probabilistically, prioritizing hyperparameter configurations that are more likely to improve performance. This targeted exploration substantially reduces the number of model evaluations relative to a traditional grid search, which exhaustively tests every combination in a predefined space.

A limitation of the proposed framework is that the vulnerability estimates and SHAP values are reported as point estimates without confidence intervals. The regularization and model-selection bias introduced by the ML estimator of the conditional mean, combined with the generated nature of the vulnerability status produced by the Monte Carlo simulation and classification steps, render standard inference methods inapplicable. A computationally feasible alternative would be to adapt the Warp-Speed bootstrap of Giacomini et al. (2013), which reduces the cost of bootstrap inference from $B \times K$ to K full-procedure evaluations by drawing a single bootstrap resample per Monte Carlo sample. Establishing the formal validity of this approach, however, requires proving bootstrap consistency for the vulnerability estimator, which is non-trivial given the multiple layers of estimation involved. We therefore leave the development of inference procedures for this framework to future work.

We now illustrate how our proposal can be used to analyze poverty vulnerability and to inform targeted, place-based public policies, applying it to the case of Ecuador.

4 An Application to Ecuador

Ecuador is a particularly apt setting in which to illustrate the framework. Its territory spans three sharply distinct geographic regions—the Pacific coast, the Andean highlands, and the Amazon basin—whose climatic, ecological, and socioeconomic conditions differ markedly, leaving the country highly but unevenly exposed to climate risk (Fernandez et al., 2015; Portalanza et al., 2024). This exposure bears directly on household welfare, since agricultural production, on which a large share of the rural population depends, is highly sensitive to temperature and precipitation variability across the country's provinces (Lopez et al., 2021; Serrano-Vincenti et al., 2025). Recent micro-level evidence confirms that such climatic fluctuations translate into poverty: rainfall shocks reduce per-capita income in rural Ecuador, with losses concentrated among poor, agricultural, and informal-sector households (Llerena Pinto et al., 2025). These pressures compound already high poverty and pronounced regional inequality, longstanding concerns for Ecuadorian policy (World Bank, 2025). Existing assessments of climate vulnerability in the country, however, rely largely on parametric or index-based methods that neither exploit the predictive structure of the simulation problem nor isolate the contribution of individual shocks across regions—precisely the dimensions our framework is designed to address.

We construct a panel for the period 2007–2021 that combines climate and household-level data. Temperature and precipitation data are drawn from CRU-TS 4.06, the gridded time series of the Climatic Research Unit of the University of East Anglia (Harris et al., 2020), downscaled with WorldClim 2.1 (Fick and Hijmans, 2017). These are historical monthly series at a spatial resolution of approximately 1 km^2, which we aggregate to the parish level—the finest geographic identifier available in the household survey.

We consider four climate shocks: maximum temperature, minimum temperature, floods, and droughts. Maximum and minimum temperatures correspond to the most extreme temperatures recorded in a given parish during a given year. Floods and droughts are constructed from the three-month Standardised Precipitation Index (SPI-3), which measures the deficit or surplus of precipitation accumulated over three months. Following McKee et al. (1993), a drought event begins when the SPI-3 falls below -1 and ends when the index returns to a positive value, with its magnitude given by the sum of the SPI-3 over the event. Symmetrically, a flood event begins when the SPI-3 rises above 1 and ends when the index returns to a negative value, with its magnitude given by the sum of the SPI-3 over the event.

We merge the climate data with the National Survey of Employment, Unemployment, and Underemployment (Encuesta Nacional de Empleo, Desempleo y Subempleo, ENEMDU), conducted by INEC in Ecuador. Our dependent variable is per-capita household income, the welfare aggregate used by INEC to determine poverty status. Because per-capita household income is constant within households, using all members would generate observations with identical outcomes but potentially different vulnerability classifications, while averaging member characteristics to the household level would conflate heterogeneity—averaging age across a head and a young dependent produces a value representing neither, and averaging binary variables such as informality loses their discrete interpretation. We therefore restrict the sample to household heads, ensuring a one-to-one mapping between observations and the household-level outcome. The head's characteristics serve as informative predictors of household welfare, since they correlate with other members' profiles through assortative matching and shared socioeconomic conditions. For each head, we observe rural residence, informality and self-employment status, sex, primary-sector employment (agriculture, hunting, forestry, fishing, or mining and quarrying), education level, and age. Merging the survey and climate data yields a data set of approximately 188,000 household-level observations covering every province in Ecuador except Galápagos, Santo Domingo, and Santa Elena—206 parishes in 21 provinces in total.

We implement the four-step procedure as follows. The first step estimates, by XGBoost regression, the equation relating the welfare aggregate to household characteristics (including year and province dummies) and climate shocks, yielding estimates of the conditional mean function and the ECDF of the residuals. Hyperparameters are selected via k-fold cross-validation: we split the data set into training and test sets, define a parameter grid over the learning rate, maximum tree depth, number of estimators, and column sampling by tree, and evaluate combinations using two folds. The grid search retains the parameter set with the best average out-of-fold performance, thereby reducing the risk of overfitting.

The second step simulates the future welfare aggregate $M = 5{,}000$ times by drawing climate and idiosyncratic shocks from their respective empirical cumulative distribution functions, the former at the province level and the latter at the national level. That is, climate shocks are drawn from the ECDF within the province to which each household belongs,[6] while idiosyncratic shocks are drawn from the ECDF for the whole sample.

[6]The stratified sampling for the climate shocks aims to simulate "reasonable" climate shocks for every household. Thus, we only consider climate shocks that have occurred in the household's province. Even though we have information at the parish level, we stratify at the province level, which serves two purposes. First, it increases the variability of simulated scenarios by allowing each parish to experience shocks observed elsewhere in its province. Second, it mitigates the concern that parishes not having experienced a particular shock during 2007–2021 would be assumed to never face such a shock, since the province-level pool encompasses a broader range of historical events.

The third step estimates household vulnerability from the simulated future income. Finally, the fourth step estimates four variants of equation (11): the regional vulnerability rate, summary statistics of household characteristics, feature importance, and the effect of each characteristic and climate shock on predicted poverty vulnerability. For the latter two objects we compute the SHAP values from a classification model in which the dependent variable is the estimated vulnerability status and the regressors are household characteristics and climate shocks.

4.1 Main results

Poverty vulnerability is concentrated in the Amazonian region of Ecuador. Figure 1 reports vulnerability rates by province in 2021. Four of the five most vulnerable provinces lie in the Amazon—Napo (0.274), Pastaza (0.179), Morona Santiago (0.141), and Sucumbíos (0.127)—pointing to deep structural disparities in this area. The only exception is Esmeraldas (0.202) on the Pacific coast, which ranks second overall. These patterns identify the eastern region as a priority for place-based poverty interventions, though Esmeraldas also warrants attention.

Figure 1: Estimated Vulnerability Rates by Province, 2021

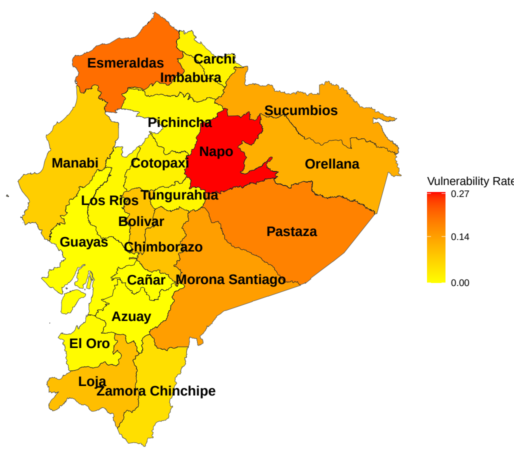
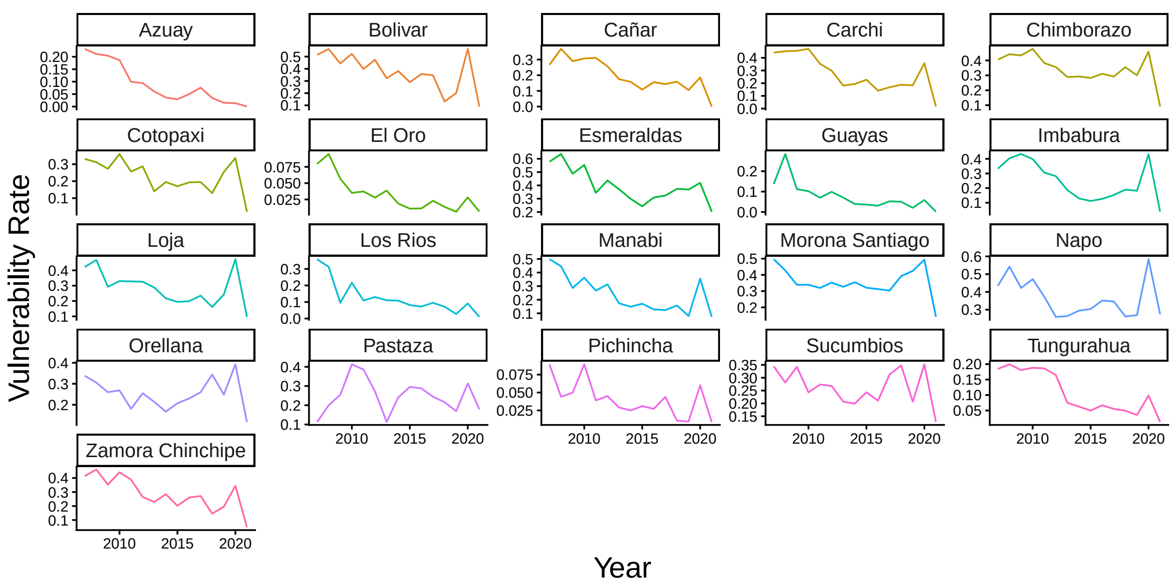


Table 1 and Figure 2 summarize the regional dynamics of poverty vulnerability from 2007 to 2021. The geographic concentration of vulnerability is highly persistent: Bolívar, in the Andean highlands, ranked first or second in most years through 2017 but improved markedly thereafter, falling to rank 9 by 2021, whereas Esmeraldas, on the Pacific coast, remained persistently near the top throughout, never falling below rank 7. By contrast, El Oro, Guayas, and Los Ríos—all coastal—consistently rank among the least vulnerable provinces, pointing to the structural advantages of coastal economies in buffering households against poverty risk. Over the study period, then, poverty vulnerability is concentrated in the Amazon region, though several Andean provinces (notably Bolívar and Chimborazo) also exhibit persistently high vulnerability. Figure 2 traces the evolution of vulnerability by province, showing a broad decline over the period, interrupted by an increase between 2019 and 2020 that reversed in 2021. Having outlined the spatial and temporal dimensions of vulnerability, we now turn to the socioeconomic profile of those most at risk. Vulnerability is concentrated among household heads who

Table 1: Rank of Provinces According to Poverty Vulnerability (Descending), 2007–21

	Bolívar	Esmeraldas	Chimborazo	Morona Santiago	Napo	Pastaza	Zamora Chinchipe	Orellana	Loja	Carchi	Sucumbios	Manabi	Cotopaxi	Imbabura	Cañar	Los Rios	Tungurahua	Guayas	Azuay	Pichincha	El Oro
2007	2	1	9	4	6	19	8	12	7	5	11	3	14	13	15	10	17	18	16	20	21
2008	2	1	8	9	3	18	5	14	4	6	16	7	13	10	11	12	19	15	17	21	20
2009	3	1	4	9	6	15	7	14	10	2	8	12	13	5	11	19	17	18	16	21	20
2010	2	1	3	11	4	7	6	14	12	5	15	9	10	8	13	16	17	19	18	20	21
2011	1	7	4	9	5	3	2	16	8	6	12	13	14	11	10	17	15	19	18	20	21
2012	1	2	3	4	13	10	12	15	5	7	11	6	8	9	14	17	16	18	19	20	21
2013	3	1	4	2	6	15	7	8	5	11	9	13	14	10	12	16	17	18	19	21	20
2014	1	3	5	2	4	7	6	12	8	11	9	14	10	15	13	16	17	18	19	20	21
2015	4	7	5	1	2	3	10	9	11	8	6	12	13	14	15	16	17	18	20	19	21
2016	1	5	4	3	2	6	7	8	10	13	9	14	11	15	12	16	17	19	18	20	21
2017	1	3	6	5	2	9	7	8	10	12	4	15	11	13	14	16	18	19	17	20	21
2018	14	2	3	1	6	7	13	5	10	9	4	12	15	8	11	16	18	17	19	21	20
2019	9	2	3	1	4	13	10	6	7	11	8	15	5	12	14	17	16	18	19	20	21
2020	2	7	5	3	1	14	12	8	4	9	11	10	13	6	15	17	16	19	21	18	20
2021	9	2	8	4	1	3	11	6	7	14	5	10	13	12	21	16	15	19	20	17	18
Mode	1	1	4	4	6	7	7	8	10	11	11	12	13	13	15	16	17	18	19	20	21
Min	1	1	3	1	1	3	2	5	4	2	4	3	5	5	10	10	15	15	16	17	18
Max	14	7	9	11	13	19	13	16	12	14	16	15	15	15	21	19	19	19	21	21	21

Figure 2: Dynamics of Poverty Vulnerability by Province, 2007–21



are low-educated, informal, self-employed, working in the primary sector, and residing in rural areas, with no systematic differences by average age, gender, or marital status. Table 2 reports average characteristics for the vulnerable and for the entire population, isolating compositional effects by comparing the two averages for each variable.[7] The gaps are striking: the share employed informally is 0.94 among the vulnerable versus

[7]If the vast majority of the sample are employed informally, finding that the vulnerable are mostly employed informally might result from the composition of the labor market.

Table 2: Average Values for household characteristics, Vulnerable vs. Entire Sample

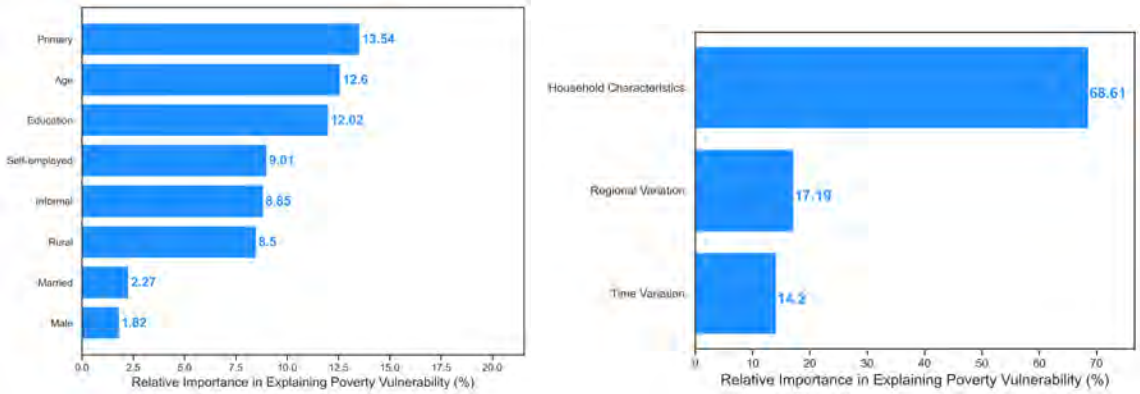
Characteristic	Vulnerable	Overall
Primary	0.83	0.32
Rural	0.86	0.39
Self-employed	0.70	0.40
Informal	0.94	0.58
Education	5.20	8.94
Married	0.51	0.48
Age	43.26	44.49
Male	0.80	0.79

0.58 in the overall sample, in the primary sector 0.83 versus 0.32, and self-employed 0.70 versus 0.40. Vulnerable households are also far more likely to reside in rural areas (0.86 vs. 0.39) and have substantially lower educational attainment (5.20 vs. 8.94 years). Average age, gender, and marital status, by contrast, show no meaningful differences between the two groups.

Poverty vulnerability is a multidimensional phenomenon, with no single characteristic dominating its prediction. Figure 3a reports the variable importance of household characteristics, computed by averaging the SHAP values across all households in absolute value and normalizing so that the shares sum to 100 (Rodríguez-Pérez and Bajorath, 2019). The six most salient characteristics—working in the primary sector (13.5), age (12.6), education level (12.0), self-employment status (9.0), informal employment (8.8), and living in rural areas (8.5)—contribute similarly, with no variable accounting for a disproportionate share of explanatory power. Gender and marital status, by contrast, play a negligible role (2.3 and 1.8, respectively). Consistent with Table 2, this pattern reinforces the multidimensional nature of vulnerability and the need for comprehensive policy interventions. Age is a notable exception: it ranks second in importance even though Table 2 shows no meaningful difference in average age between the vulnerable and the overall population (43.26 vs. 44.49). By capturing heterogeneous effects at the individual level, the SHAP values reveal that age is in fact a key predictor of vulnerability whose influence is simply obscured by averaging.

Figure 3b complements this picture by decomposing explanatory power into its aggregate components. Household characteristics alone account for nearly 70 percent of the explained variation in vulnerability, confirming that socioeconomic conditions are the primary predictor of poverty risk. Regional and time variation account for 17 and 14 percent, respectively; despite aggregating 21 province dummies and 15 year dummies, these components together explain substantially less than the household-level characteristics.

Figure 3: Variable Importance of household characteristics in Explaining Vulnerability



(a) Household characteristics (b) Aggregate Relative Importance

The importance of explaining poverty vulnerability in figure 3 is computed according to the average SHAP values (in absolute value) and normalized so that the sum of variable importance adds up to 100.

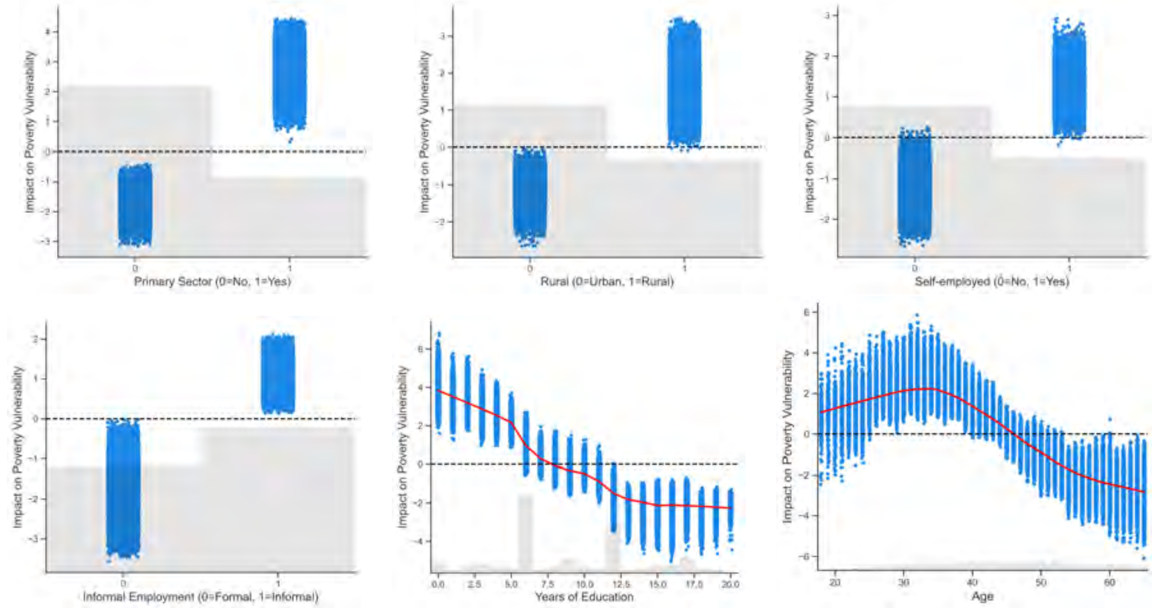
Figure 4 summarizes the impact of the most relevant variables on poverty vulnerability. Each dot depicts the SHAP value of the corresponding variable for a household in the sample, making plain that the effect of each characteristic on the probability of vulnerability is heterogeneous across households: were informality, say, to affect vulnerability homogeneously, all points would lie at a single value along the y-axis. For education and age, the effect is not only heterogeneous at a given value but also nonlinear.

The results in Figure 4 for the binary variables align with Table 2: working in the primary sector, living in a rural area, and being self-employed or an informal worker all raise the probability of being vulnerable. While the direction of these effects is consistent across households—the SHAP values are uniformly positive or negative for these characteristics—the spread of points along the y-axis indicates that their magnitude differs considerably, pointing to meaningful heterogeneity in how each characteristic shapes poverty risk.

Education is negatively associated with vulnerability. The relationship is strongest at low levels of schooling, where moving from zero to five years of education corresponds to the largest decline in predicted vulnerability, and it declines monotonically thereafter, with higher education remaining protective throughout albeit at a more moderate slope. Notably, every household head with secondary or higher education exhibits a negative SHAP value, so that higher education is associated with a lower predicted probability of vulnerability for every head in the sample.

The effect of age is nonlinear, which explains why Table 2 reports no systematic difference in average age between the vulnerable and the overall population: the mean

Figure 4: Effects of household characteristics on Poverty Vulnerability



Note: The impacts of each of the five variables on poverty vulnerability are represented by the SHAP values shown in the figure. The dots in each figure depict the SHAP value of the corresponding variable for a household in the sample, highlighting that the effect of each characteristic on the probability of poverty vulnerability is heterogeneous across observations. The solid LOWESS line represents a locally weighted regression that smooths the scatter of SHAP values, capturing the underlying nonlinear relationship between the variable and its marginal contribution to vulnerability across the distribution.

comparison simply averages out effects of opposite sign. SHAP values rise at younger ages, peak around age 34, and then decline monotonically, crossing zero around age 44. Between roughly ages 25 and 37, all households display positive SHAP values, so that heads in this range are uniformly associated with higher vulnerability; beyond age 50 the pattern reverses just as clearly, with nearly all SHAP values turning negative and reflecting a strong, consistent association with lower vulnerability.

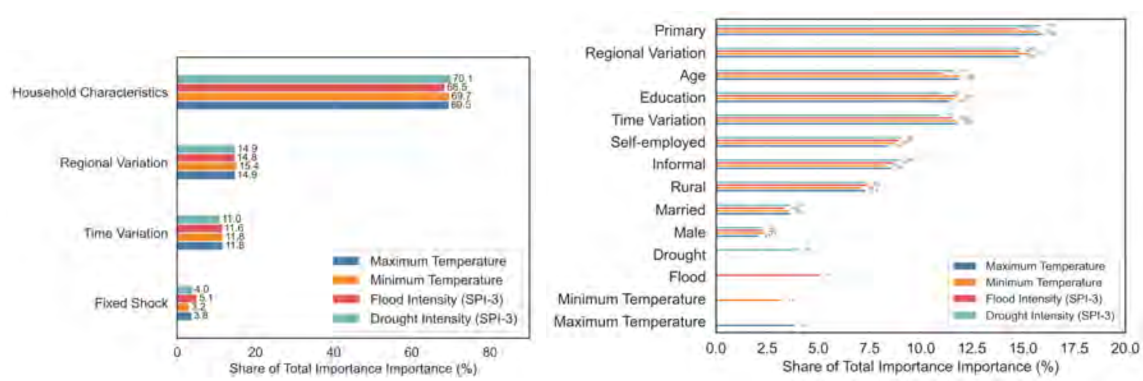
These results point to formalization as the most actionable policy for reducing poverty vulnerability, with targeted efforts prioritizing self-employed primary-sector workers of low educational attainment residing in rural areas, particularly those aged 25 to 37. Expanding access to secondary education is a complementary but longer-term strategy, given the strong, nonlinear, and protective association between schooling and vulnerability documented above, whose largest gains accrue to those with fewer than five years of education.

Having characterized the socioeconomic profile of the vulnerable, we turn to the role of climate shocks. We assess the predictive relevance of each shock and estimate its heterogeneous impact on vulnerability across regions by fixing the realized exposure to each shock while simulating the remaining ones from their joint empirical distribution, following the procedure leading to equation (15). Two limitations are worth noting.

First, our simulations draw climate shocks from the historical distribution observed during 2007–2021; if climate change increases the frequency or severity of extreme events, as the growing evidence on fat tails suggests, our estimates may understate future risk and should be read as lower bounds on climate vulnerability. Incorporating forward-looking climate projections, such as those from Ecuador's Country Climate and Development Report (CCDR) or regional climate models, would help address this and represents a natural extension for future work. Second, because each exercise fixes a single shock, the framework does not permit an assessment of the joint contribution of all climate shocks; fixing all shocks simultaneously would leave only idiosyncratic variation in the simulated welfare aggregate. Developing an approach to evaluate their joint relevance is therefore a further promising direction for future work.

Figure 5 reports the relative importance of climate shocks for poverty vulnerability. Panel (a) gives the aggregated contribution of household characteristics, regional variation, time variation, and the fixed climate shock. Across all shock-specific exercises, household characteristics account for the largest share of explanatory power (around 70 percent), reinforcing that vulnerability is primarily explained by structural socioeconomic conditions, while regional and time variation contribute more modest shares of around 15 and 12 percent, respectively. As panel (b) shows, the individual variable importances are likewise stable across exercises. Both the aggregated and the individual shares thus remain stable across all shock-specific specifications and closely track those reported in Figure 3b, reinforcing the robustness of the decomposition.

Figure 5: Relative Importance of Household, Regional, Temporal, and Climate Factors in Explaining Vulnerability



(a) Aggregated (b) Individual

The importance of explaining poverty vulnerability is computed according to the average SHAP values (in absolute value) and normalized so that the sum of variable importance adds up to 100.

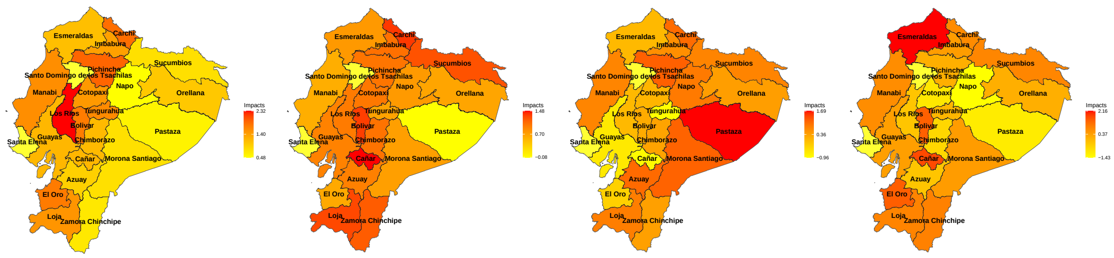
Although household characteristics explain the largest share, climate shocks remain a relevant predictor of poverty vulnerability. Regional and temporal variation are captured by 21 province dummies and 15 year dummies, respectively, whereas the fixed climate shock is a single variable; yet its contribution, ranging from 3.2 to 5.1 percent across exercises, is notable both relative to these extensive sets of fixed effects and relative to individual household characteristics. For reference, Figure 3b shows that self-employment status, informal employment, and rural residence—three of the most salient household-level predictors—account for 9, 8.8, and 8.5 percent, respectively, so that the contribution of floods (5.1 percent) and droughts (4.0 percent) is not dramatically smaller. Moreover, unlike employment status or rural residence, which are persistent features of a household's socioeconomic position, climate shocks are transitory events affecting all households in a region. While secondary to structural socioeconomic factors in the aggregate decomposition, climate shocks are therefore a quantitatively meaningful and policy-relevant predictor of poverty vulnerability.

Targeting climate adaptation policy requires identifying not only which shocks are most predictive of vulnerability but also where their effects concentrate. Because climate shocks are by nature extreme events, averaging their effect across the entire distribution of realized values would obscure their impact. To focus on the most relevant cases, we identify the four most affected regions by computing, for each province separately, the average SHAP value for flood and drought realizations at or above the 90th percentile of that province's shock distribution; for minimum and maximum temperatures, given their smaller variation, we average the SHAP value over the households experiencing the four most extreme temperatures.

Figure 6 shows the spatial variation in the impact of each climate shock on vulnerability across provinces. Floods display a localized pattern, with the strongest effects concentrated in the north-central provinces, particularly Los Ríos, Pichincha, and Carchí. Droughts follow a distinct geography, with impacts most pronounced in the central Andean provinces—especially Cañar—and extending across the southern highlands. Maximum temperature shocks bite hardest in the Amazonian region, suggesting that households there are particularly sensitive to extreme heat, whereas minimum temperature shocks matter most in the north and center, in provinces such as Esmeraldas, Cañar, and Bolívar. These contrasts confirm that the impacts of different shocks vary geographically, underscoring the need for spatially differentiated adaptation and mitigation strategies.

Figure 7 depicts the heterogeneous impact of flood events in the four most affected provinces: Los Ríos, Pichincha, Carchí, and El Oro. The underlying trend is remarkably homogeneous across provinces—negative or negligible effects at low flood values, crossing zero at moderate intensities, and rising steeply at the upper tail. Yet the

Figure 6: Impacts of Specific Climate Shocks on Vulnerability by Province



(a) Floods (b) Droughts (c) Maximum temperature (d) Minimum temperature

Note: Each panel represents the average SHAP value for shock realizations at or above the 90th percentile (in case of temperature, four most extreme temperatures) of the province-specific shock distribution, focusing on the most extreme realizations to avoid diluting the signal with moderate ones.

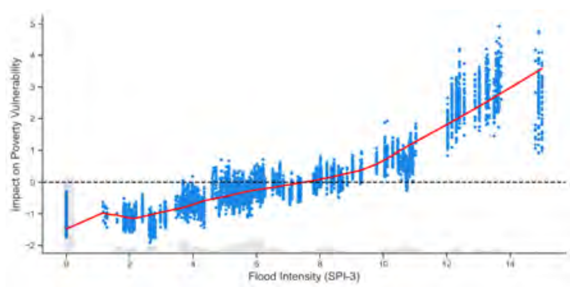
dispersion of individual SHAP values around the curve is wide throughout, indicating that, conditional on flood intensity, households within the same province experience very different impacts on their vulnerability. This within-province heterogeneity echoes the individual-level differences documented in Figure 4, and suggests that adaptation policies will be more effective when they combine regional targeting with the socioeconomic profile of the most vulnerable.

Figure 8 turns to drought shocks in the four most exposed provinces. A broadly common pattern emerges—mild drought conditions carry negligible or negative effects, while more severe rainfall deficits are associated with increasingly positive impacts—but drought effects are markedly more heterogeneous across provinces than floods, with substantial dispersion of individual effects at every drought intensity and notable differences in how the trend evolves from province to province.

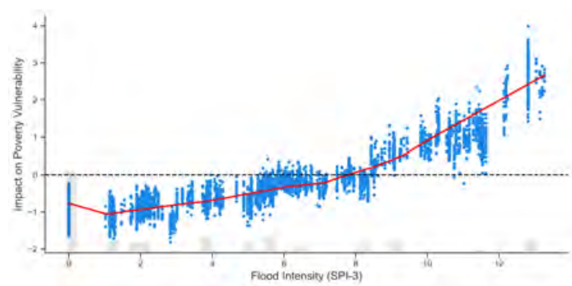
Figure 9 examines maximum temperature shocks. Pastaza and Morona Santiago display a clear threshold: effects are negligible or negative at moderate temperatures and turn sharply positive only at the upper tail. Azuay follows an asymmetric pattern, with elevated vulnerability at both temperature extremes and reduced risk at intermediate values around 22–23°C, though the decline from low temperatures is sharper than the subsequent increase. Pichincha shows a small cluster of observations at very low temperatures with positive effects; its highest concentration of observations sits at moderate temperatures, where risk is low, while high temperatures again carry a positive association with vulnerability.

Figure 10 displays the SHAP values of minimum temperature by region, illustrating how extreme increases in minimum temperature affect predicted vulnerability while household characteristics are held constant and the remaining shocks are simulated. For

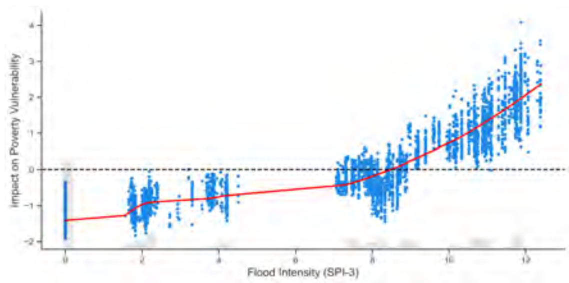
Figure 7: Heterogeneous Impact of Floods on Poverty Vulnerability in the Four Most Affected Provinces



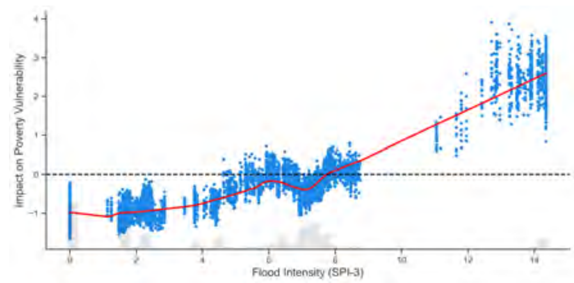
(a) Los Ríos



(b) Pichincha



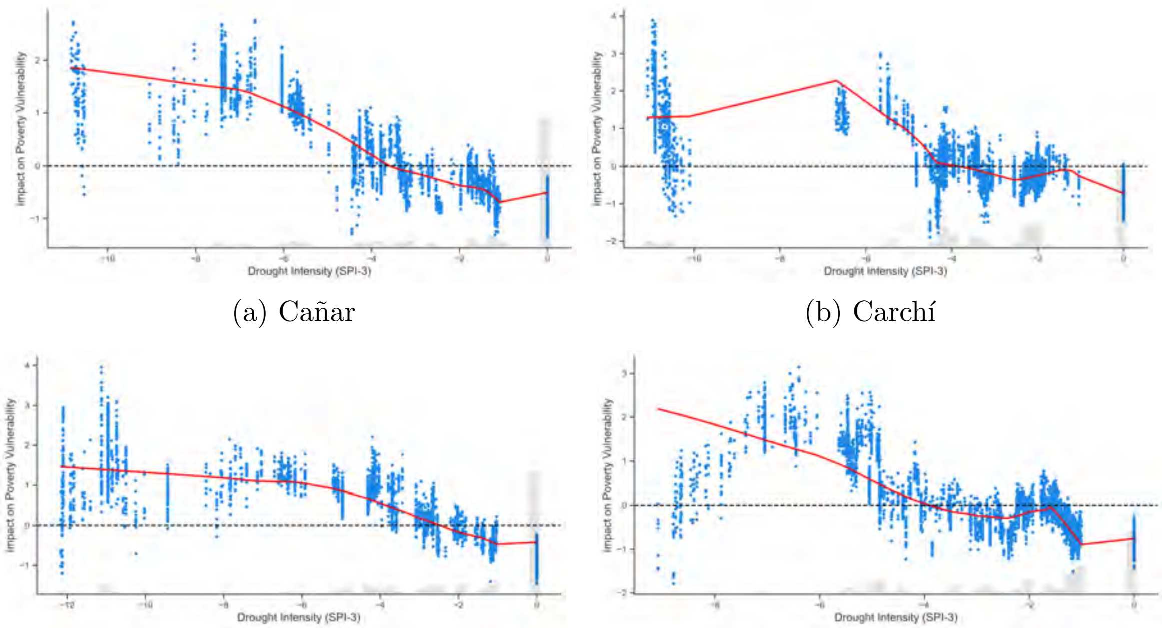
(c) Carchí



(d) El Oro

Note: This figure represents the SHAP values associated with the fixed shock for a household in the sample. The solid LOWESS line represents a locally weighted regression that smooths the scatter of SHAP values, capturing the underlying nonlinear relationship between the variable and its marginal contribution to vulnerability across the distribution. The four provinces displayed were selected because they rank among the regions with the highest average marginal contributions to vulnerability using the 90th percentile.

Figure 8: Heterogeneous Impact of Droughts on Poverty Vulnerability in the Four Most Affected Provinces



(a) Cañar

(b) Carchí

(c) Bolívar

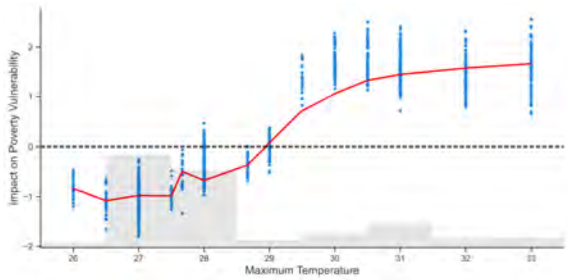
(d) Loja

Note: This figure represents the SHAP values associated with the fixed shock for a household in the sample. The solid LOWESS line represents a locally weighted regression that smooths the scatter of SHAP values, capturing the underlying nonlinear relationship between the variable and its marginal contribution to vulnerability across the distribution. The four provinces displayed were selected because they rank among the regions with the highest average marginal contributions to vulnerability using the 90th percentile.

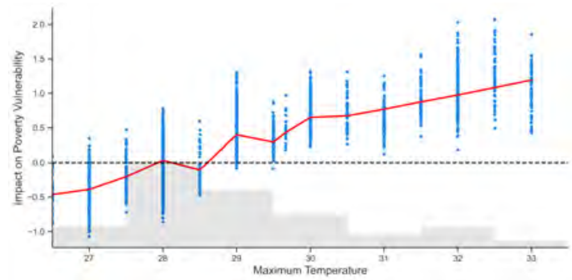
Esmeraldas, Cañar, and El Oro the pattern is consistent: impacts are positive at very low minimum temperatures and decline toward zero (or turn negative) as temperatures rise to moderate levels, indicating that cold extremes are positively associated with poverty risk. Bolívar presents the opposite pattern, with impacts that are negative at low temperatures and increase as minimum temperatures rise. Across all four provinces, the nonlinear and heterogeneous nature of these relationships underscores the limits of one-size-fits-all responses and the value of region-specific interventions tailored to local climatic conditions.

Although climate shocks account for a modest share of aggregate vulnerability, the preceding analysis shows that their effects are spatially concentrated and can be severe locally. Particular regions exhibit sharp sensitivity to floods, droughts, or extreme heat even when national averages remain limited, and within any given region the impact of a shock varies considerably across households, reflecting differences in labor market status, sector of employment, educational attainment, and rural residence. The policy implication is that climate-related poverty risks in Ecuador are best addressed through interventions that are both geographically targeted and directed to the socioeconomic

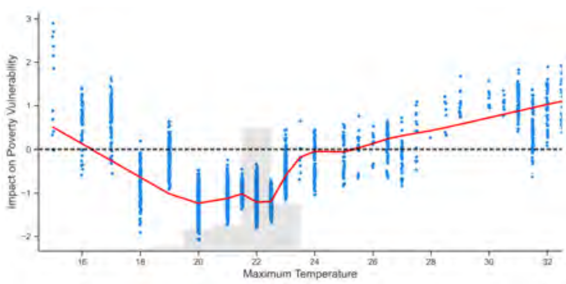
Figure 9: Heterogeneous Impact of Maximum Temperature on Poverty Vulnerability in the Four Most Affected Provinces



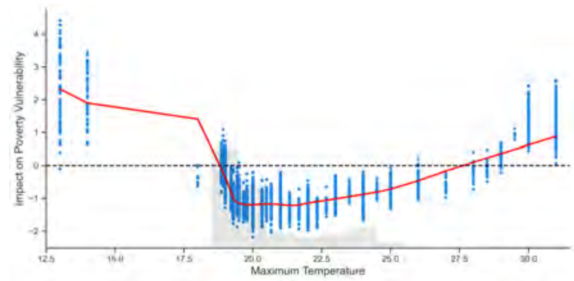
(a) Pastaza



(b) Morona Santiago



(c) Azuay



(d) Pichincha

Note: This figure represents the SHAP values associated with the fixed shock for a household in the sample. The solid LOWESS line represents a locally weighted regression that smooths the scatter of SHAP values, capturing the underlying nonlinear relationship between the variable and its marginal contribution to vulnerability across the distribution. The four provinces displayed were selected because they rank among the regions with the highest average marginal contributions to vulnerability using the four most extreme temperatures.

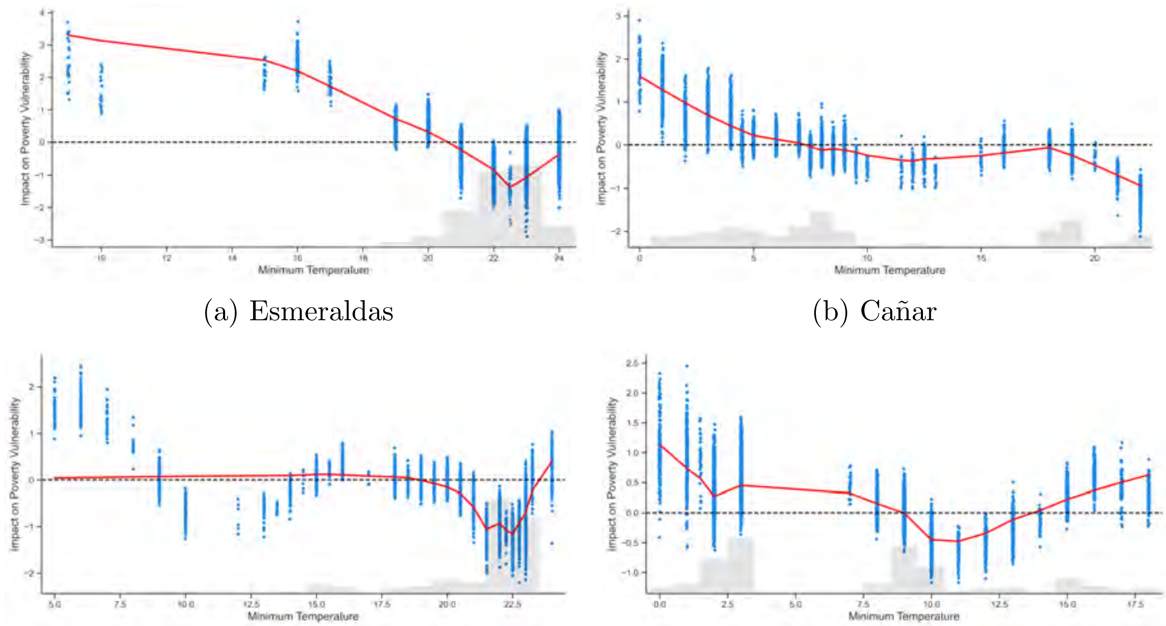
profile of the most exposed, rather than uniform national or even regional responses. In drought-prone Andean provinces, investments in water management, irrigation infrastructure, and agricultural support are likely to help; in flood-exposed north-central zones, localized prevention and drainage maintenance are more cost-effective than broad transfer programs. Across all shock types, the results point to the value of region-specific social protection instruments that can be scaled up rapidly following extreme events, rather than non-targeted programs that dilute resources across unaffected areas.

4.2 Robustness checks

4.2.1 OLS estimation

To motivate the use of XGBoost, we repeat the vulnerability analysis using a standard OLS specification from the original HPA framework. Comparing the two sets of results tells us whether a flexible, nonlinear approach is necessary to capture the relationships among household characteristics, climate shocks, and vulnerability. We estimate the conditional mean of income given household characteristics and climate shocks with the

Figure 10: SHAP Values of Minimum Temperature by Region



(a) Esmeraldas (b) Cañar

(c) El Oro (d) Bolivar

Note: This figure represents the SHAP values associated with the fixed shock for a household in the sample. The solid LOWESS line represents a locally weighted regression that smooths the scatter of SHAP values, capturing the underlying nonlinear relationship between the variable and its marginal contribution to vulnerability across the distribution. The four provinces displayed were selected because they rank among the regions with the highest average marginal contributions to vulnerability using the four most extreme temperatures.

specification

$$y_{irt} = \alpha + \beta' \mathbf{S}_{rt} + f(Age_{irt}) + g(Edu_{irt}) + \theta' \mathbf{X}_{irt} + \phi' (\mathbf{S}_{rt} \times \mathbf{X}_{irt}) + \mu_r + \lambda_t + \varepsilon_{irt},$$

where y_{irt} denotes the household income of individual i in region r and year t; $\mathbf{S}_{rt}$ is a vector of climate shocks at the region-year level; $\mathbf{X}_{irt}$ is a vector of individual and household characteristics; $\mathbf{S}_{rt} \times \mathbf{X}_{irt}$ captures the interaction effects between shocks and characteristics; $f(Age_{irt})$ and $g(Edu_{irt})$ denote polynomial (quadratic) functions of age and years of education; μ_r and λ_t are region and year fixed effects; and ε_{irt} is an idiosyncratic error term. As in the nonparametric case, we simulate future income using the empirical distribution of shocks and residuals, and compute household vulnerability as the probability that simulated income falls below the poverty line.

Table B1 reports the average household characteristics for the vulnerable population and the overall sample under the OLS-based simulation. In stark contrast to Table 2, virtually no meaningful differences emerge between the two groups: the shares of rural residents, primary-sector workers, informal employees, and self-employed are nearly identical across the vulnerable and the overall population. This is telling. By imposing a linear structure on a potentially highly nonlinear relationship among household characteristics, climate shocks, and welfare, OLS fails to discriminate between vulnerable and non-vulnerable households; and the large number of interactions needed to capture heterogeneous shock effects under a linear specification introduces overfitting that distorts the predicted welfare aggregate and, in turn, the estimated vulnerability status.

The OLS vulnerability map in Figure B1 differs substantially from its XGBoost counterpart in Figure 1: many provinces receive a vulnerability rate of zero, and the ranking of the most vulnerable provinces changes dramatically, with Azuay, Guayas, and El Oro—precisely the provinces one would least expect at the top—displacing the Amazonian provinces that dominate under XGBoost. Rather than capturing actual geographic variation in poverty risk, this reflects a misspecified conditional mean that smooths over cross-province differences and overlooks the localized nonlinear effects of climate shocks that XGBoost recovers. Taken together, these findings point to the advantages of the machine learning approach and the limitations imposed by the more restrictive assumptions of OLS. The failure of OLS to produce plausible vulnerability classifications also makes the SHAP-based characterization uninformative under this specification, since there is little to learn from decomposing a measure that cannot distinguish vulnerable from non-vulnerable households.

4.2.2 Full Sample

As an additional robustness check, we repeat the analysis on the full sample of approximately 400,000 individuals rather than the roughly 188,000 household heads used in the main specification, allowing us to assess whether the results are sensitive to the unit of observation. The findings confirm the robustness of the main conclusions: while the magnitude of vulnerability rates and the dispersion of SHAP values differ slightly with the change in sample composition, the spatial patterns, the socioeconomic profile of the vulnerable, and the heterogeneous effects of climate shocks all remain remarkably stable.

The spatial and temporal patterns of vulnerability remain largely consistent with the main specification. Figure B2 shows provincial vulnerability in 2021, closely matching Figure 1, with the Amazonian provinces—particularly Napo, Pastaza, and Morona Santiago—together with Esmeraldas remaining the most exposed; vulnerability rates are somewhat smoother because of the presence of multiple household members, but the relative ranking is preserved. The dynamics over 2007–2021 (Figure B3) replicate the overall decline in Figure 2, the temporary increase around 2020, and the subsequent reduction in 2021, with the same provinces persistently among the most vulnerable.

Average characteristics (Table B2) and provincial rankings (Table B3) closely match those in Tables 2 and 1, again concentrating vulnerability among rural, less-educated, primary-sector, informal, and self-employed individuals. Minor shifts appear in intermediate ranks and in some characteristics, but the modal ranks of the most vulnerable provinces remain stable, and the provinces at the top of the ranking in the household-head sample also lead in the full sample.

Figures B4 and B5 replicate the variable-importance and SHAP-value plots for the full sample, and the results align closely with the main specification. The aggregate decomposition in panel (b) of Figure B4 is nearly identical to its counterpart in Figure 3: household characteristics account for roughly 68 percent of the explained variation, while regional and time variation contribute 18 and 14 percent, respectively. At the disaggregated level the ranking of individual predictors is likewise preserved—primary-sector employment (15.3), education (12.0), informality (10.3), and rural residence (10.1) remain the dominant contributors, with gender and marital status negligible in both specifications. The main difference is that self-employment loses importance relative to the household-head sample (3.8 versus 9.0 percent), likely because this characteristic is less uniformly predictive once non-heads are included. The SHAP profiles in Figure B5 confirm that the qualitative relationships are unchanged: primary-sector employment, informality, self-employment, and rural residence all raise vulnerability; education exerts a strong, nonlinear protective effect that is largest at low schooling levels and diminishes but persists thereafter; and the age profile retains its shape, peaking around the early thirties and declining steadily beyond the mid-forties. The consistency of these nonlinear patterns across sample definitions confirms that they are not an artifact of restricting the analysis to household heads.

Figure B6 confirms that the aggregate decomposition is highly stable across sample definitions. Panel (a) closely mirrors Figure 5: across all four shock-specific specifications, household characteristics account for between 68.5 and 69.6 percent of the explained variation, regional variation for between 16.2 and 16.8 percent, and time variation for between 11.2 and 11.4 percent—all within one to two percentage points of their household-head counterparts. The contribution of the fixed climate shock is similarly stable, ranging from 2.9 to 3.5 percent across exercises, somewhat below the 3.2 to 5.1 percent range in the main specification, though the ordering across shocks is preserved, with floods contributing the most and maximum temperature the least. Panel (b) reinforces this conclusion: primary-sector employment remains the single most important household-level predictor (around 16 to 17 percent), followed closely by regional variation, education, informality, age, and rural residence, all within one to two percentage points of the shares in Figure 5. The main difference is that self-employment and the head-of-household indicator together absorb a share previously concentrated in self-employment alone, which mechanically reduces the individual contribution of self-employment. In short, the results are not driven by the unit of observation.

Figure B7 replicates the provincial shock-impact maps for the full sample, and the geographic patterns are broadly consistent with Figure 6: flood impacts remain concentrated in the north-central areas, drought effects in the central Andean highlands and south, maximum temperature shocks along the Amazonian region, and minimum temperature shocks on the northern coast. Intensity levels differ slightly with sample composition, but the geographic clustering of shock sensitivity is preserved.

Figures B8, B9, B11, and B10 replicate the regional SHAP profiles for floods, droughts, and maximum and minimum temperature under the full sample, and the key nonlinear patterns of the main specification remain clearly visible. For floods, the increasing relationship between intensity and vulnerability persists, with Los Ríos, Carchí, and El Oro still among the most affected; Bolívar replaces Pichincha as the fourth most exposed province, though the shape of the LOWESS curves is otherwise unchanged. For droughts, the profiles confirm the predominantly positive, monotonically increasing relationship between rainfall deficits and vulnerability, with Bolívar and Loja appearing in both specifications and the remaining provinces shifting modestly without altering the overall pattern. For maximum temperature, the threshold pattern documented in Azuay, Pastaza, and Pichincha—negligible effects at moderate temperatures turning sharply positive in the upper tail—is preserved, with Carchí replacing Morona Santiago as the second most exposed province. For minimum temperature, the pattern described for Esmeraldas, Cañar, and El Oro remains the same, with Loja replacing Bolívar as the fourth most affected region and exhibiting a very similar profile.

Taken together, the full-sample results confirm that the spatial patterns, the socioeconomic profile of the vulnerable, and the heterogeneous effects of climate shocks documented in the main specification are robust to expanding the sample beyond household heads to all household members.

5 Discussion and Conclusion

In this paper we formalize and generalize a simulation approach to poverty vulnerability. We first show that it can be framed as an out-of-sample prediction problem, which motivates estimating the conditional mean of welfare with regularized machine learning rather than OLS. We then generalize the procedure from vulnerability measures to functions of vulnerability and show how these functions, recovered through SHAP values, inform the design of targeted, place-based policy.

Framing vulnerability estimation as prediction is what makes the machine-learning approach appropriate. As Kleinberg et al. (2015) argue, an important class of policy problems turns on predictive rather than causal inference, and for these the minimization of out-of-sample prediction error—the objective of machine-learning methods—is the relevant criterion, one to which unbiased coefficient estimation is not tuned. Because the second step of the framework evaluates the estimated conditional mean at shock realizations not used in estimation, it is precisely such a prediction problem. Regularized, flexible learners are well suited to it, they accommodate the nonlinearities and interactions that a linear specification imposes away, while guarding against overfitting through regularization and cross-validation (Mullainathan and Spiess, 2017; Athey and Imbens, 2019). Our OLS robustness check bears this out, as the linear specification fails to separate vulnerable from non-vulnerable households once the requisite interactions are introduced.

A second benefit is interpretability. A recurring objection to flexible learners is that predictive gains come at the cost of opacity, which is a serious limitation when the goal is to target policy. By attributing each prediction to its inputs at the observation level, SHAP values (Lundberg, 2017) let us move from a black-box predictor to a description of who is vulnerable and of how specific climate shocks alter that vulnerability across regions, turning the model into an instrument for characterization rather than prediction alone.

A third benefit is the framework's modest data requirements. Estimating vulnerability is difficult precisely because it concerns a future, counterfactual distribution of welfare, and long household panels that would identify it directly are rarely available in developing countries (Dercon, 2002; Chaudhuri et al., 2002). By simulating that distribution from a single cross-section or a short panel merged with historical climate data, the framework sidesteps this constraint and remains implementable where data are most scarce.

These features together speak to the design of place-based policy. Combining survey data with auxiliary information to obtain welfare estimates at fine spatial resolution is the logic of poverty mapping (Elbers et al., 2003), and our framework extends it to a forward-looking, shock-specific object by identifying not only where vulnerability concentrates but also which shocks drive it and whom they affect. The application to Ecuador yields four concrete findings in this spirit. First, poverty vulnerability is concentrated in the Amazonian region among low-educated, informal, self-employed, primary-sector workers residing in rural areas, particularly those aged 25 to 37. Second, vulnerability is multidimensional, no single characteristic dominates the profile, and household-level conditions jointly account for roughly 70 percent of the explained variation. This points to formalization as the most actionable lever, directed at self-employed primary-sector workers with low educational attainment in rural areas. Third, although climate shocks are secondary in the aggregate decomposition, their effects are spatially concentrated and can be severe at the local level, floods bite hardest in north-central coastal and Andean provinces, droughts in the central Andean highlands, and extreme heat in the Amazon. Within any given province, moreover, the impact of a shock varies considerably across households, reflecting differences in labor-market status, sector of employment, educational attainment, and rural residence. Fourth, taken together, these findings imply that effective policy must combine structural interventions addressing the socioeconomic drivers of vulnerability with regionally targeted instruments designed around the specific shocks each area faces, rather than relying on uniform national responses. Because the framework identifies both where and to whom shocks matter, it is well suited to informing adaptive social protection systems that can be scaled up rapidly following extreme events (Bowen et al., 2020).

The framework relies on historical climate distributions to simulate future shocks and may therefore understate vulnerability if extreme events become more frequent or severe under climate change; in this sense, the estimates should be read as lower bounds. Integrating forward-looking climate projections from established modeling exercises would sharpen the estimates and align the framework with existing climate-risk assessments (Hallegatte et al., 2017).

More broadly, the proposed framework is easy to implement and, crucially, applicable to short panels and cross-sectional data—often the only sources available in developing-country contexts. Through the empirical application to Ecuador, we show how it generates actionable, spatially and socially disaggregated insights for identifying vulnerable regions and population groups, thereby supporting the design of more precise, context-specific policies to reduce poverty vulnerability, including that arising from climate shocks.

References

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Bowen, T., del Ninno, C., Andrews, C., Coll-Black, S., Gentilini, U., Johnson, K., Kawasoe, Y., Kryeziu, A., Maher, B., and Williams, A. (2020). *Adaptive Social Protection: Building Resilience to Shocks*. International Development in Focus. World Bank, Washington, DC.

Calvo, C. and Dercon, S. (2013). Vulnerability to individual and aggregate poverty. *Social Choice and Welfare*, 41(4):721–740.

Canavire-Bacarreza, G., Conconi, A., Olivieri, S., and Serio, M. (2026). Understanding vulnerability to poverty and natural disasters in latin america and the caribbean. *Advances in Econometrics (Forthcomming)*.

Canavire-Bacarreza, G. and Serio, M. (2026). Vulnerability to poverty, climate change, and natural hazards in ecuador. *Journal of Environmental Economics and Policy (Forthcomming)*.

Chaudhuri, S. (2000). Empirical methods for assessing household vulnerability to poverty. *School of International and Public Affairs, Columbia University, New York*.

Chaudhuri, S., Jalan, J., and Suryahadi, A. (2002). Assessing household vulnerability to poverty from cross-sectional data: A methodology and estimates from indonesia. Technical Report 0102-52, Columbia University Department of Economics.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

Cred, U. (2020). Human cost of disasters. an overview of the last 20 years: 2000–2019. Technical report, CRED, UNDRR, Geneva.

Dercon, S. (2002). Income risk, coping strategies, and safety nets. *The World Bank Research Observer*, 17(2):141–166.

Đekić, J. P., Mitković, P. B., Dinić-Branković, M. M., Igić, M. Z., Đekić, P. S., and Mitković, M. P. (2018). The study of effects of greenery on temperature reduction in urban areas. *Thermal Science*, 22(Suppl. 4):988–1000.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1):355–364.

Fernandez, M. A., Bucaram, S. J., and Renteria, W. (2015). Assessing local vulnerability to climate change in Ecuador. *SpringerPlus*, 4(1):738.

Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

Giacomini, R., Politis, D. N., and White, H. (2013). A warp-speed method for conducting monte carlo experiments involving bootstrap estimators. *Econometric Theory*, 29(3):567–589.

Günther, I. and Harttgen, K. (2009). Estimating households vulnerability to idiosyncratic and covariate shocks: A novel method applied in madagascar. *World Development*, 37(7):1222–1234.

Hallegatte, S., Bangalore, M., Bonzanigo, L., Fay, M., Kane, T., Narloch, U., Rozenberg, J., Treguer, D., and Vogt-Schilb, A. (2016). *Shock Waves: Managing the Impacts of Climate Change on Poverty.* Climate Change and Development. World Bank, Washington, DC.

Hallegatte, S., Vogt-Schilb, A., Bangalore, M., and Rozenberg, J. (2017). *Unbreakable: Building the Resilience of the Poor in the Face of Natural Disasters*. Climate Change and Development. World Bank, Washington, DC.

Harris, I., Osborn, T. J., Jones, P., and Lister, D. (2020). Version 4 of the cru ts monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, 7(1):109.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12.

Hill, R. V. and Porter, C. (2017). Vulnerability to drought and food price shocks: Evidence from ethiopia. *World Development*, 96:65–77.

INEC (2023). Encuesta de condiciones de vida. Technical report, INEC.

Kleemann, J., Koo, H., Hensen, I., Mendieta-Leiva, G., , Kahnt, B., Kurze, C., Inclán, D. J., Cuenca, P., Noh, J. K., Hoffmann, M. H., Lenert, M., Lozano, P., and Fürst, C. (2022). Priorities of action and research for the protection of biodiversity and ecosystem services in continental ecuador. *Biological Conservation*, 265:109404.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–495.

Ligon, E. and Schechter, L. (2003). Measuring vulnerability. *The Economic Journal*, 113(486):C95–C102.

Llerena Pinto, M. C., Mirzabaev, A., and Qaim, M. (2025). Effects of recurrent rainfall shocks on poverty and income distribution in rural Ecuador. *World Development*, 195.

Lopez, G., Gaiser, T., Ewert, F., and Srivastava, A. (2021). Effects of recent climate change on maize yield in Southwest Ecuador. *Atmosphere*, 12(3):299.

Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Macintyre, H. L. and Heaviside, C. (2019). Potential benefits of cool roofs in reducing heat-related mortality during heatwaves in a european city. *Environment International*, 127:430–441.

McKee, T. B., Doesken, N. J., and Kleist, J. (1993). The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume 17, pages 179–183. Boston.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Portalanza, D., Torres, M., Rosso, F., Zuluaga, C. F., Durigon, A., Horgan, F. G., Alava, E., and Ferraz, S. (2024). Climate variability and change in Ecuador: Dynamic downscaling of regional projections with RegCM4 and HadGEM2-ES for informed adaptation strategies. *Frontiers in Climate*, 6:1344868.

Pritchett, L., Suryahadi, A., and Sumarto, S. (2000). Quantifying vulnerability to poverty: A proposed measure, with application to indonesia. Technical report, SMERU Research Institute.

Rodríguez-Pérez, R. and Bajorath, J. (2019). Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *Journal of Medicinal Chemistry*, 63(16):8761–8777.

Serrano-Vincenti, S., Guamán-Pozo, J., Chuqui, J., Tufiño, R., and Franco-Crespo, C. (2025). Measuring the effects of climate change on traditional crops in tropical highlands, Ecuador. *Frontiers in Sustainable Food Systems*, 9:1447593.

Shapley, L. S. (1953). A value for n-person games. *Contribution to the Theory of Games*, 2.

Shcherbatyi, I., Louppe, G., and Head, T. (2024). Scikit-learn hyperparameter search wrapper. https://scikit-optimize.github.io/stable/auto_examples/sklearn-gridsearchcv-replacement.html. Accessed: 2025-02-27.

Skoufias, E. and Quisumbing, A. R. (2003). Consumption insurance and vulnerability to poverty. Technical report, International Food Policy Research Institute (IFPRI).

Williams, C. K. I. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*, volume 2. MIT Press, Cambridge, MA.

World Bank (2014). Global facility for disaster reduction and recovery.

World Bank (2025). Rekindling the reduction of poverty and inequality in Ecuador. Technical report, World Bank, Washington, DC.

World Meteorological Organization (2021). Wmo atlas of mortality and economic losses from weather, climate and water extremes (1970–2019). Technical Report WMO-No. 1267, World Meteorological Organization, Geneva.

Appendix

A Conditional Mean Characterization for Tree and Kernel Methods

Consider again the decomposition of the welfare aggregate into observed and unobserved components,

$$C = m(\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\eta}) + \epsilon, \quad \mathbb{E}[\epsilon|\boldsymbol{X}, \boldsymbol{S}] = 0,$$

where $\boldsymbol{X}$ and $\boldsymbol{S}$ denote household characteristics and climate shocks, respectively; $m(\cdot) \coloneqq \mathbb{E}[C|\boldsymbol{X}, \boldsymbol{S}]$ is the conditional mean function, known up to the nuisance parameter $\boldsymbol{\eta}$; and ϵ is an idiosyncratic shock capturing the unexplained component. Under linearity of $m(\cdot)$, the nuisance parameter is simply the vector of regression coefficients. Because our proposal estimates $m(\cdot)$ nonparametrically, we now characterize it for two leading cases: tree-based algorithms and kernel machines.

Trees partition the covariate space into J regions $\boldsymbol{R} \coloneqq (R_1, ..., R_J)$.[8] For a fixed partition $\boldsymbol{R}$, the region coefficients $\boldsymbol{C} \coloneqq (C_1, ..., C_J)$ solve the OLS problem

$$\arg\min_{c} \frac{1}{n} \sum_{i=1}^{n} (Y_i - m(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}))^2, \quad m(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}) = \sum_{j=1}^{J} c_j 1(\boldsymbol{X}_i, \boldsymbol{S}_i \in R_j),$$

on the training set,[9] whose solution is $\hat{\boldsymbol{C}} = (\hat{c}_1, ..., \hat{c}_J)$, with $\hat{c}_j = \overline{Y}_j$ the average outcome in region R_j. Hence, for tree-based algorithms,

$$m(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}) = \sum_{j=1}^{J} c_j 1(\boldsymbol{X}_i, \boldsymbol{S}_i \in R_j),$$

and $\boldsymbol{\eta} = (\boldsymbol{C}, \boldsymbol{R})$. Any ensemble of trees—bagging, boosting, or random forests—can be used to estimate $m(\cdot)$.

Kernel machines, in contrast, solve

$$\arg\min_{m \in \mathcal{H}, ||m||_{\mathcal{H}} \leq k} \frac{1}{n} \sum_{i=1}^{n} (Y_i - m(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}))^2,$$

[8] A partition of a set is a grouping of its elements into nonempty subsets such that every element belongs to exactly one subset. For example, $\mathcal{P} = \{[0, 0.5], [0.5, 1]\}$ is a partition of $[0, 1]$.

[9] Machine learning algorithms split the data into a training set, used to fit the model, and a test set, used to evaluate it.

on the training set, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS). To each such $\mathcal{H}$ corresponds a unique positive semidefinite symmetric kernel $K(x_1, s_1; x_2, s_2)$ admitting the representation

$$K(x_1, s_1; x_2, s_2) = \sum_{j=0}^{\infty} \alpha_j \, \phi_j(x_1, s_1) \, \phi_j(x_2, s_2)$$

for a positive sequence α_j and linearly independent functions ϕ_j, such that every element $m \in \mathcal{H}$ takes the form

$$m(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}) = \sum_{j=0}^{\infty} m_j \phi_j(\boldsymbol{X}_i, \boldsymbol{S}_i), \quad \sum_{j=0}^{\infty} \frac{m_j}{\alpha_j} < \infty.$$

The kernel-machine problem admits the closed-form solution

$$\hat{m}(x, s) = \sum_{i=1}^{n} \hat{\alpha}_i \, K(x, s; \boldsymbol{X}_i, \boldsymbol{S}_i), \qquad \hat{\boldsymbol{\alpha}} = (\boldsymbol{K} + \lambda_n I_n)^{-1} \boldsymbol{Y}, \tag{A1}$$

where $\boldsymbol{K}$ is the $n \times n$ matrix with ij-th element $K(\boldsymbol{X}_i, \boldsymbol{S}_i; \boldsymbol{X}_j, \boldsymbol{S}_j)$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ is the vector of outcomes, and $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)'$ collects the estimated coefficients.[10]

A common choice is the squared-exponential kernel (Williams and Rasmussen, 2006),

$$K(x_1, x_2) = \sigma_f^2 \times \exp\left\{-\frac{|x_1 - x_2|^2}{2l^2}\right\},$$

for which

$$m(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}) = \sum_{j=0}^{\infty} m_j \phi_j(\boldsymbol{X}_i, \boldsymbol{S}_i),$$

and $\boldsymbol{\eta} = \left(\lambda_n, \sigma_f^2, l^2\right)$.

[10] Equation (A1) shows that kernel machines generalize ridge regression.

B Robustness Checks

Table B1: Average Values for household characteristics, Vulnerable vs. Entire Sample (OLS)

Characteristic	Vulnerable	Overall
Rural	0.35	0.39
Primary	0.29	0.32
Married	0.51	0.48
Male	0.77	0.79
Education	9.09	8.94
Self-employed	0.40	0.40
Age	44.85	44.49
Informal	0.57	0.58

Figure B1: Estimated Vulnerability Rates by Province, 2021 (OLS)

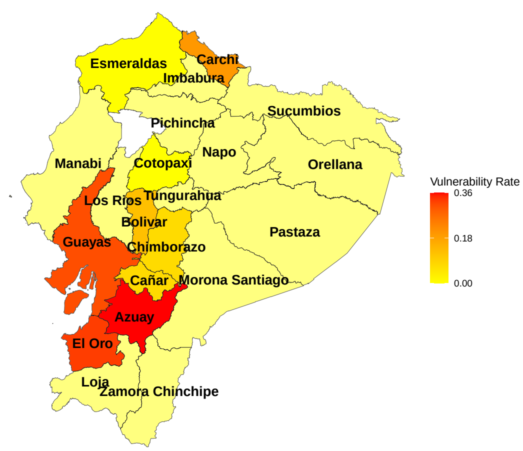


Figure B2: Estimated Vulnerability Rates by Province, 2021 (Full Sample)

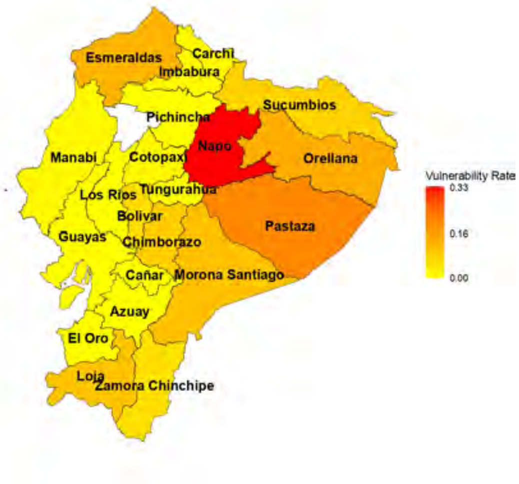


Table B2: Average Values for household characteristics, Vulnerable vs. Entire Sample (Full Sample)

Characteristic	Vulnerable	Overall
Primary	0.89	0.31
Rural	0.91	0.40
Informal	0.96	0.61
Education	5.64	9.57
Self-employed	0.43	0.32
Married	0.45	0.41
Head	0.49	0.47
Age	38.37	39.29
Male	0.57	0.58

Table B3: Rank of Provinces Using the Full Sample According to Poverty Vulnerability (Descending), 2007–21

	Bolívar	Napo	Morona Santiago	Pastaza	Chimborazo	Esmeraldas	Zamora Chinchipe	Carchi	Loja	Orellana	Sucumbios	Cotopaxi	Cañar	Imbabura	Manabi	Azuay	Tungurahua	Los Rios	Guayas	Pichincha	El Oro
2007	1	6	2	18	10	5	4	8	7	9	11	13	15	12	3	16	17	14	19	21	20
2008	3	2	8	19	6	1	7	4	5	11	14	13	10	9	12	16	18	15	17	21	20
2009	1	2	3	5	4	6	9	7	10	12	16	11	13	8	15	14	17	20	18	21	19
2010	1	3	7	5	2	4	11	8	6	13	16	9	12	10	14	15	17	18	19	20	21
2011	1	5	7	10	4	14	3	8	2	12	13	11	9	6	16	17	15	19	18	20	21
2012	1	9	2	8	5	6	7	14	3	4	10	13	11	12	15	17	16	18	19	21	20
2013	2	6	1	7	5	4	11	14	3	10	9	13	8	12	15	16	17	18	19	20	21
2014	1	3	2	4	5	7	6	12	8	9	11	10	13	14	15	18	17	16	19	20	21
2015	4	1	2	3	6	8	10	12	9	5	7	11	13	15	14	18	17	16	19	20	21
2016	2	1	5	3	4	9	7	13	8	6	10	11	12	14	15	17	16	18	20	19	21
2017	1	2	4	5	3	7	8	11	10	6	9	12	13	14	15	18	17	16	19	20	21
2018	13	6	2	9	1	7	12	8	5	3	4	11	10	14	15	19	16	17	18	20	21
2019	9	5	1	7	3	2	11	12	8	4	10	6	13	14	15	17	16	19	18	20	21
2020	2	1	3	13	6	7	14	11	4	5	9	12	15	10	8	21	17	16	18	19	20
2021	9	1	5	2	7	4	10	13	6	3	8	11	16	12	14	20	15	21	18	19	17
Mode	1	1	2	5	6	7	7	8	8	9	10	11	13	14	15	17	17	18	19	20	21
Min	1	1	1	2	1	1	3	4	2	3	4	6	8	6	3	14	15	14	17	19	17
Max	13	9	8	19	10	14	14	14	10	13	16	13	16	15	16	21	18	21	20	21	21

Figure B3: Dynamics of Poverty Vulnerability by Province, 2007–21

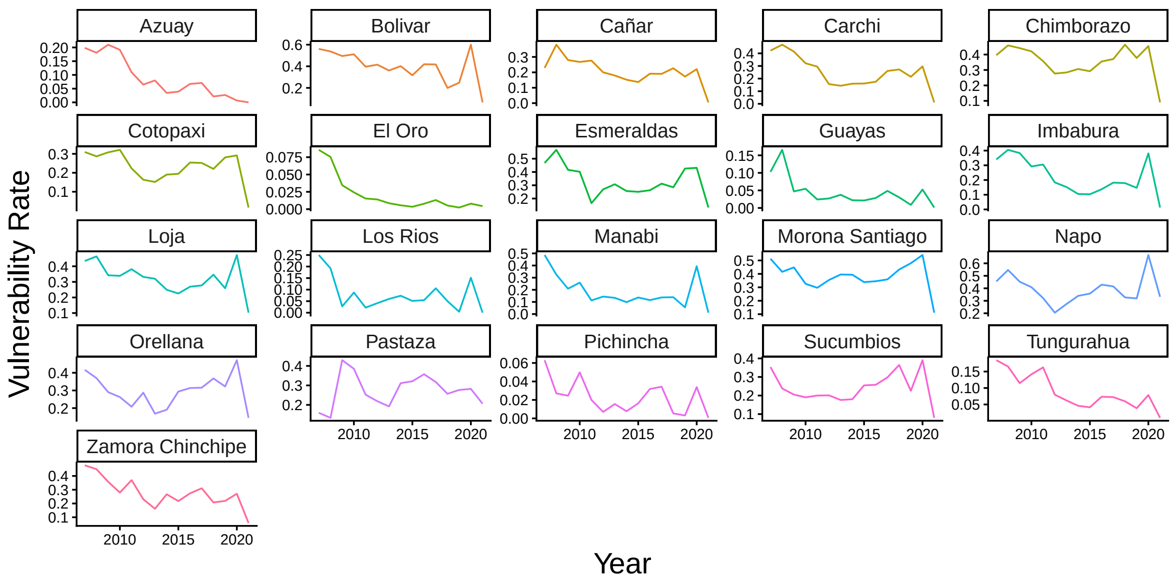
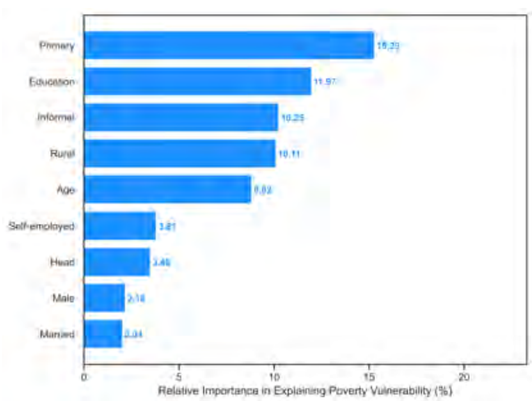
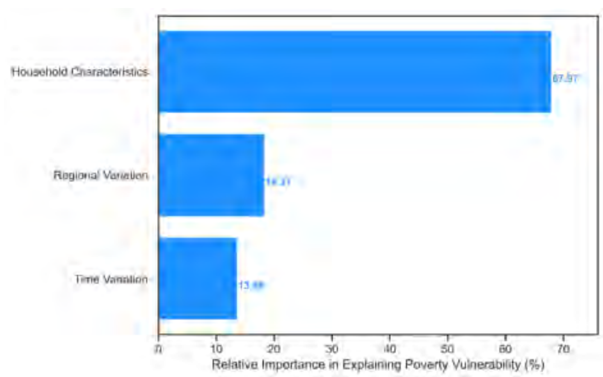


Figure B4: Variable Importance of household characteristics in Explaining Vulnerability (Full Sample)



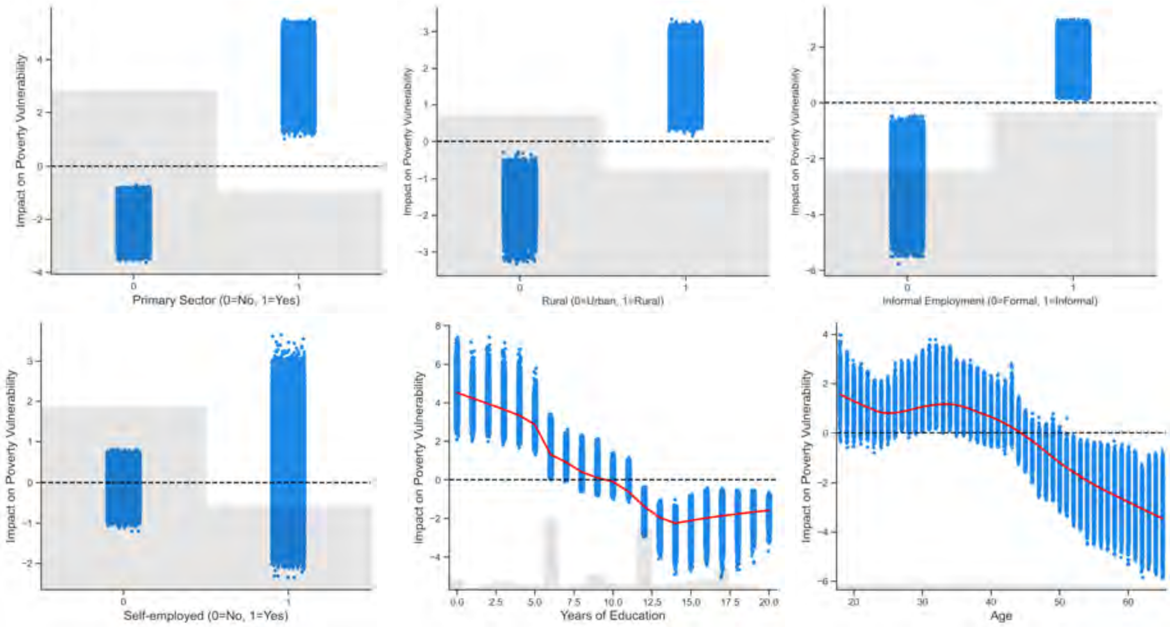
(a) Household characteristics



(b) Aggregate Relative Importance

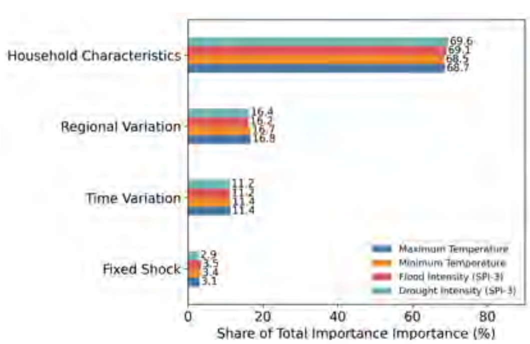
The importance of explaining poverty vulnerability in figure B4 is computed according to the average SHAP values (in absolute value) and normalized so that the sum of variable importance adds up to 100.

Figure B5: Effects of household characteristics on Poverty Vulnerability (Full Sample)

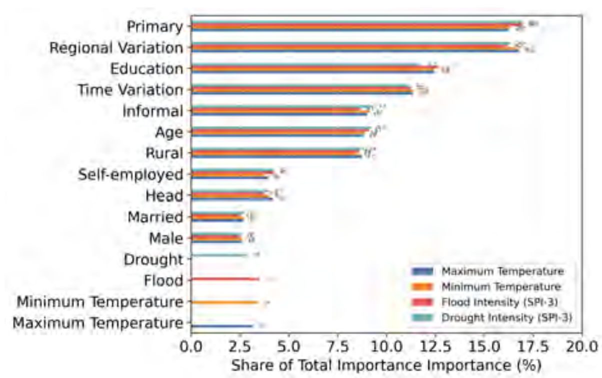


Note: The impacts of each of the five variables on poverty vulnerability are represented by the SHAP values shown in the figure. The dots in each figure depict the SHAP value of the corresponding variable for a household in the sample, highlighting that the effect of each characteristic on the probability of poverty vulnerability is heterogeneous across observations. The solid LOWESS line represents a locally weighted regression that smooths the scatter of SHAP values, capturing the underlying nonlinear relationship between the variable and its marginal contribution to vulnerability across the distribution.

Figure B6: Relative Importance of Household, Regional, Temporal, and Climate Factors in Explaining Vulnerability (Full Sample)



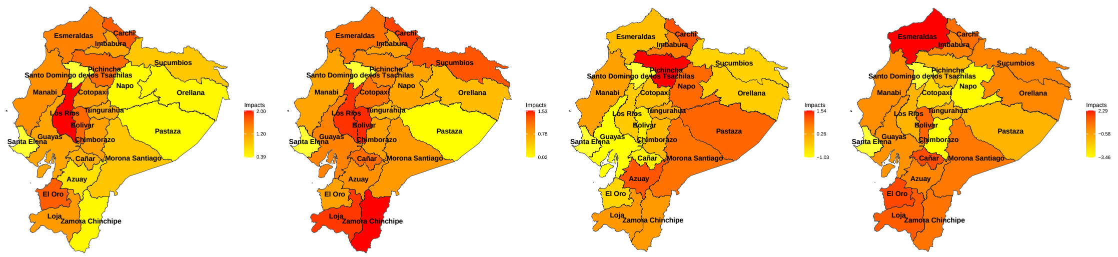
(a) Aggregated



(b) Individual

Note: The importance of specific characteristics in explaining poverty vulnerability presented in figure B6 is computed according to the average SHAP values (in absolute value), and normalized so that the sum of variable importance adds up to 100.

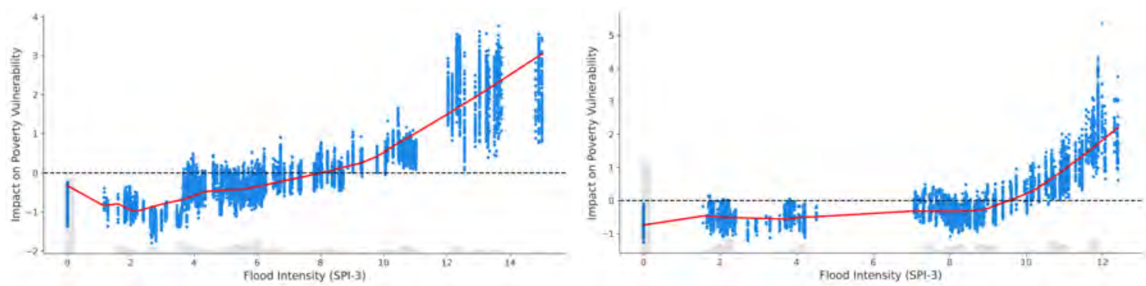
Figure B7: Impacts of Specific Climate Shocks on Vulnerability by Province (Full Sample)



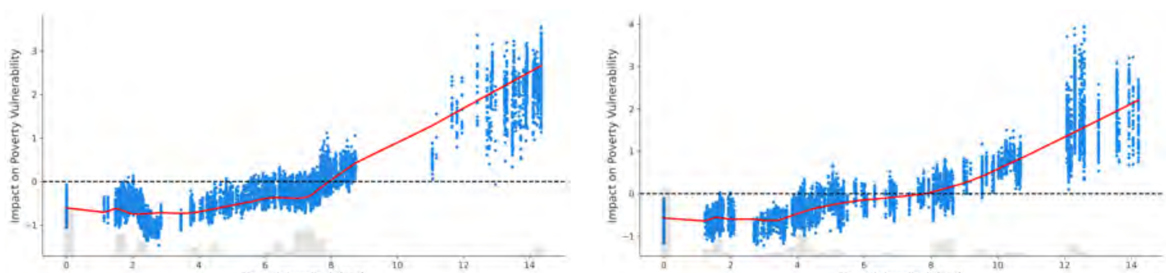
(a) Floods (b) Droughts (c) Maximum temperature (d) Minimum temperature

Note: Each panel represents the average SHAP value for households that experienced a shock realization at or above the 90th percentile (in case of temperature, four most extreme temperatures) of the province-specific shock distribution, focusing on the most extreme realizations to avoid diluting the signal with moderate ones.

Figure B8: Heterogeneous Impact of Floods on Poverty Vulnerability in the Four Most Affected Provinces (Full Sample)



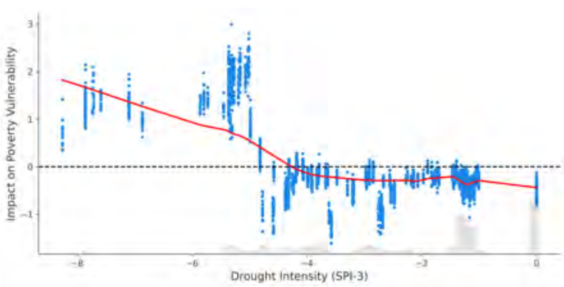
(a) Los Ríos (b) Carchí



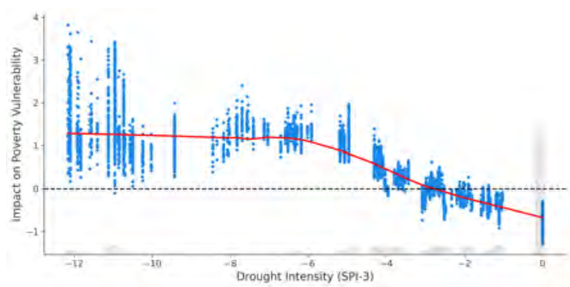
(c) El Oro (d) Bolívar

Note: This figure represents the SHAP values associated with the fixed shock for a household in the sample. The solid LOWESS line represents a locally weighted regression that smooths the scatter of SHAP values, capturing the underlying nonlinear relationship between the variable and its marginal contribution to vulnerability across the distribution. The four provinces displayed were selected because they rank among the regions with the highest average marginal contributions to vulnerability using the 90th percentile.

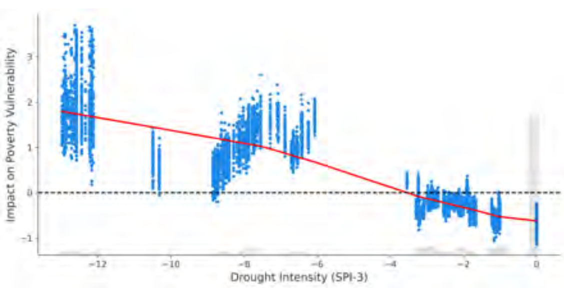
Figure B9: Heterogeneous Impact of Droughts on Poverty Vulnerability in the Four Most Affected Provinces (Full Sample)



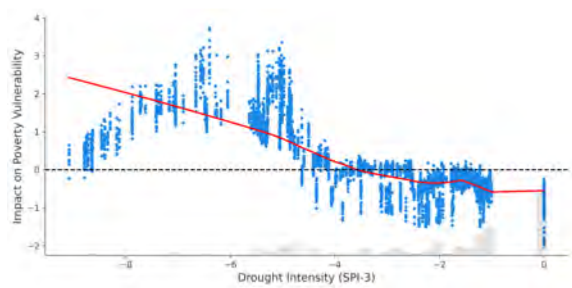
(a) Zamora Chinchipe



(b) Bolívar



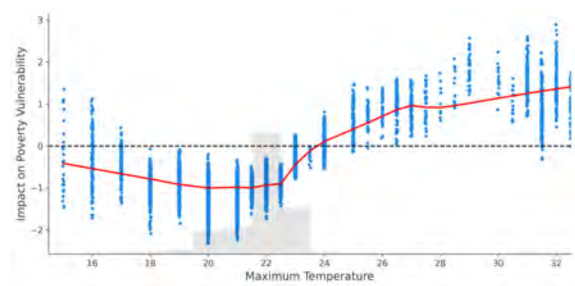
(c) Los Ríos



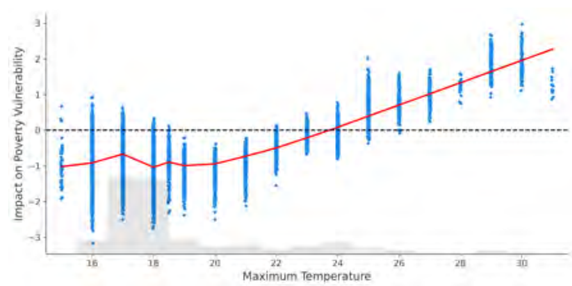
(d) Loja

Note: This figure represent the SHAP values associated with the fixed shock for a household in the sample. The solid LOWESS line represents a locally weighted regression that smooths the scatter of SHAP values, capturing the underlying nonlinear relationship between the variable and its marginal contribution to vulnerability across the distribution. The four provinces displayed were selected because they rank among the regions with the highest average marginal contributions to vulnerability using the 90th percentile.

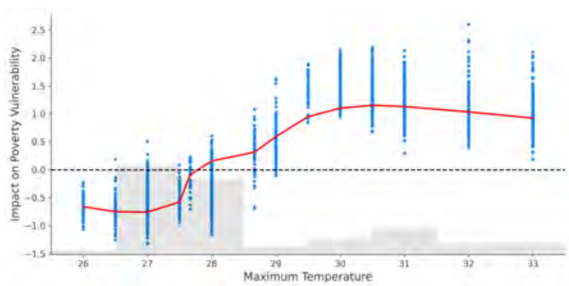
Figure B10: Heterogeneous Impact of Maximum Temperature on Poverty Vulnerability in the Four Most Affected Provinces (Full Sample)



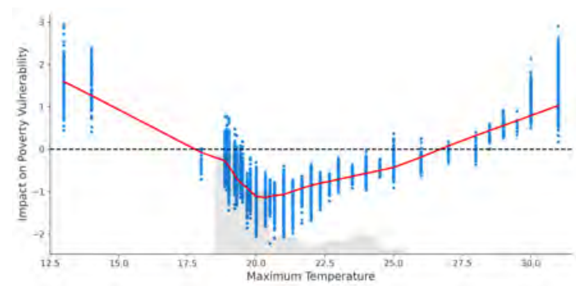
(a) Azuay



(b) Carchi



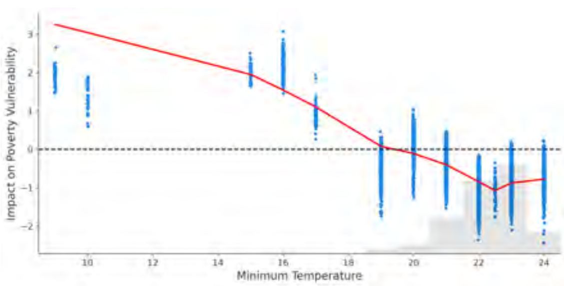
(c) Pastaza



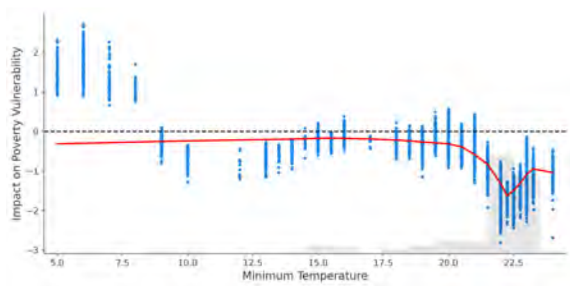
(d) Pichincha

Note: This figure represent the SHAP values associated with the fixed shock for a household in the sample. The solid LOWESS line represents a locally weighted regression that smooths the scatter of SHAP values, capturing the underlying nonlinear relationship between the variable and its marginal contribution to vulnerability across the distribution. The four provinces displayed were selected because they rank among the regions with the highest average marginal contributions to vulnerability using the four most extreme temperatures.

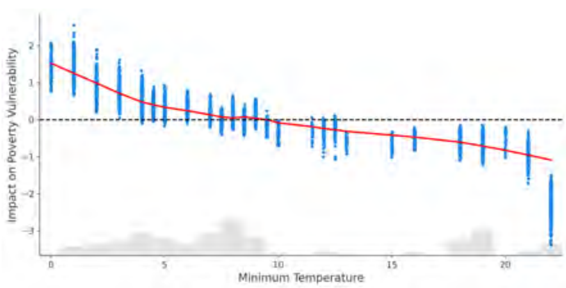
Figure B11: Heterogeneous Impact of Minimum Temperature on Poverty Vulnerability in the Four Most Affected Provinces (Full Sample)



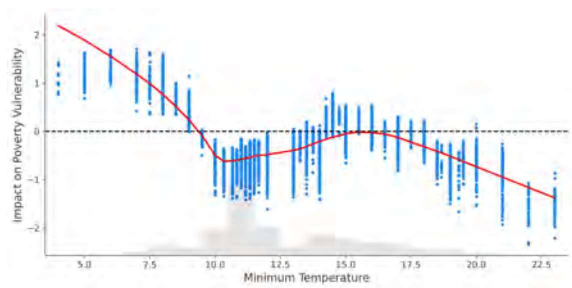
(a) Esmeraldas



(b) El Oro



(c) Cañar



(d) Loja

Note: This figure represent the SHAP values associated with the fixed shock for a household in the sample. The solid LOWESS line represents a locally weighted regression that smooths the scatter of SHAP values, capturing the underlying nonlinear relationship between the variable and its marginal contribution to vulnerability across the distribution. The four provinces displayed were selected because they rank among the regions with the highest average marginal contributions to vulnerability using the four most extreme temperatures.