

CLARE

A Causal machine Learning Approach to Resilience Estimation

Talip Kilic

Marco Letta

Pierluigi Montalbano

Federica Petruccelli



WORLD BANK GROUP

Development Economics
Development Data Group
January 2026



Reproducible Research Repository

A verified reproducibility package for this paper is available at <http://reproducibility.worldbank.org>, click [here](#) for direct access.

Abstract

This paper proposes a new resilience index, CLARE (Causal machine Learning Approach to Resilience Estimation), which is rooted in an impact evaluation framework and causal machine learning algorithms applied to longitudinal household survey data. The indicator is model-agnostic, data-driven, scalable, and normatively anchored to wellbeing thresholds, and can be either shock-specific or a general-purpose resilience metric. The paper provides an empirical demonstration of constructing the CLARE resilience index, leveraging more than 28,000 household observations from 19 nationally representative, longitudinal, multi-topic surveys that were implemented by the national statistical offices in Malawi, Nigeria, Tanzania, and Uganda over 2009–20 in partnership with the World Bank Living Standards Measurement Study. Although the paper centers on measuring resilience to drought, the proposed index is applicable to any type of shock. The analysis shows

that CLARE outperforms existing resilience metrics and alternative approaches to predict food insecurity out-of-sample—both in the future (dynamic forecasting) and in held-out countries (cross-sectional prediction). The index can be decomposed to causally identify the relative importance of resilience capacities that can insulate populations from shocks. Thus, it can be operationalized in designing, targeting, and monitoring policies and investments that aim to strengthen resilience. CLARE’s deployment—paired with continued investments in national longitudinal survey platforms—can boost the effectiveness of early-warning systems and resilience-building interventions, while allowing the transfer of resilience policy advice from data-rich contexts to data-poor environments that may not immediately provide the requisite longitudinal survey data for index estimation.

This paper is a product of the Development Data Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://www.worldbank.org/prwp>. The authors may be contacted at tkilic@worldbank.org and marco.letta@uniroma1.it. A verified reproducibility package for this paper is available at <http://reproducibility.worldbank.org>, click **here** for direct access.



The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

CLARE: A Causal machine Learning Approach to Resilience Estimation

Talip Kilic[‡], Marco Letta[§], Pierluigi Montalbano[§], Federica Petruccelli^{§1}

Authorized for distribution by Haishan Fu, Chief Statistician and Director, Development Data Group, Development Economics,
World Bank Group

JEL-Codes: C31; O12; O15.

Keywords: resilience; causal machine learning; longitudinal household surveys, impact evaluation; predictive analytics; policy targeting.

¹ ‡ Development Data Group, World Bank; tkilic@worldbank.org; § Department of Social Sciences and Economics, Sapienza University of Rome; marco.letta@uniroma1.it; pierluigi.montalbano@uniroma1.it; federica.petruccelli@uniroma1.it. We are grateful to Chris Barrett for his valuable advice and insightful comments on an earlier version of this work, and to Alemayehu Ambel, Adriana Paolantonio, Donato Romano, Luca Tiberti, Philip Wollburg, and participants at the *CSAE 2025 Conference* in Oxford, the *SITES X Development Conference* in Rome, and the *Better Data for Better Jobs and Lives: Innovations in Survey Measurement in the Age of AI* conference in Washington for their helpful comments and suggestions. We also thank Siobhan Murray for her assistance with integrating the geospatial weather data.

1. Introduction

In the era of polycrisis, identifying the subjects most in need of resilience-building interventions is crucial. Essential requirements for effectively targeting and evaluating these interventions include the scalability of resilience indicators, their explicit anchoring to the myriad of shocks and stressors affecting wellbeing, and the alignment between the adopted resilience measures and the wellbeing outcomes they are meant to reflect. In other words, resilience indicators should accurately predict wellbeing measures out of sample. By ‘out-of-sample prediction’, we refer to the ability of a resilience indicator to correctly identify wellbeing outcomes when applied to previously unseen data. These can be either data from other households, areas, or countries referred to the same point in time, or new data from future time periods. However, the available resilience measures often lack scalability, are not explicitly tied to shocks, and exhibit at best mediocre ability to predict out-of-sample these wellbeing outcomes (Upton et al., 2022). As the concept of resilience has been embraced by a growing number of donors, agencies, and international organizations, and is currently shaping the humanitarian and development agenda (Béné et al., 2017), these empirical problems translate into a significant policy obstacle to the expanding array of resilience-oriented interventions. Indeed, policy discussions increasingly emphasize that progress on climate resilience—a prominent dimension of the broader resilience agenda—is constrained not only by limited resources, but also by weak measurement systems that fail to identify who is resilient, to what shocks, and through which channels. For example, recent World Bank work stresses that tracking resilience remains limited and that better, decision-relevant metrics are essential for targeting and learning in resilience-building policies (Shilpi et al., 2025).

We address these problems by leveraging Machine Learning (ML) algorithms to construct a novel indicator of household resilience. These algorithms excel at predicting out-of-sample and were designed to address prediction policy problems (Kleinberg et al., 2015; Athey & Imbens, 2019). However, we do not simply rely on predictive ML techniques, which have now become standard tools in the econometric toolkit of empirical economists. Our approach incorporates recent innovations in causal ML (Wager & Athey, 2018; Chernozhukov et al., 2018; Chernozhukov et al., 2024) by employing state-of-the-art ML algorithms specifically adapted to predict heterogeneous causal effects rather than outcomes. The intuition behind the construction of our indicator—labeled CLARE (an acronym for a Causal machine Learning Approach to Resilience Estimation)—is straightforward. CLARE is a simple weighted average of standardized resilience subcomponents commonly identified in the specialized literature, normalized on a 0–100 scale for interpretability and rescaled by the probability of exceeding a normative wellbeing

threshold under the shock.² The novelty lies in using causal ML to construct a data-driven weighting scheme for aggregating the resilience components, derived from the counterfactual evaluation of the underlying causal relationships between wellbeing, shocks, and resilience drivers. As such, CLARE is based on a linear aggregation method using weights estimated non-linearly, thereby balancing complexity and interpretability. In this way, CLARE addresses the challenging issue of credibly aggregating different resilience subcomponents into a single composite indicator of resilience. While composite indicators offer a synthetic and comprehensive understanding of the phenomena under study, particularly with latent variables such as resilience, they also make it challenging to pinpoint the specific roles and weights of each component. The widespread and often discretionary practice—typically adopted for simplicity—of assuming equal weights or making other subjective choices during aggregation presents significant methodological and practical challenges. These include inconsistencies with the complexity of real-world data-generating processes and potential threats to the reliability of policy targeting, which frequently relies on composite indicators to guide prioritization and resource allocation.

This is where our causal ML approach comes into play: we use it to derive importance weights for each resilience subcomponent in intermediating the causal relationship between the shock of interest and the outcome to which resilience capacity is intended to be indexed. Consequently, the entire causal ML procedure serves the main purpose of conducting a preliminary counterfactual estimation of the relationship between wellbeing outcomes, covariate shocks, and intermediating variables. This estimation ultimately yields convex aggregation weights for each resilience subcomponent included in CLARE, as well as the out-of-sample probability of being above a given normative wellbeing threshold (e.g., food insecurity) under the shock. The final index is then computed at the household level as a weighted average of all subcomponents, with aggregation weights (non-negative and summing up to one) determined on causal patterns learned from the data, rescaled by the household-specific probability of being food secure under the shock. This approach bears some resemblance to the re-weighting of control units in the donor pool for constructing the counterfactual evolution of the outcome of the treated unit in the synthetic control method (Abadie et al., 2010).

² By resilience subcomponents, we refer to the elementary components of the composite resilience indicator. In our context, these represent the drivers of the conditional average treatment effects of the shock on the wellbeing outcome, ranked by importance. Therefore, this term does not refer to intermediate constructs such as resilience 'pillars' or 'capacities' found in frameworks like RIMA or TANGO, as CLARE moves directly from raw resilience components to the final composite indicator, bypassing intermediate constructs.

Importantly, while in the illustrative application we show how to compute CLARE using a specific technique—causal forests (Athey et al., 2019; Wager & Athey, 2018)—CLARE is not tied to any specific methodology. As long as the chosen approach estimates granular and heterogeneous treatment effects and enables the establishment of an objective hierarchy among the drivers of heterogeneity, any causal ML technique can be employed to identify the underlying causal relationships. Similarly, the definition of ‘variable importance’ can vary depending on the approach and the importance ranking method chosen (e.g., SHAP values, permutation feature importance). Thus, while based on the use of one method from the family of causal ML techniques, CLARE is model-agnostic.

Despite being data-driven, CLARE is not a black box. First, the identification of the underlying causal relationships is thoroughly rooted in the potential outcomes framework (Imbens & Rubin, 2015). Second, while CLARE remains agnostic regarding the relative importance of each subcomponent in dampening the impacts of shocks, the selection of subcomponents included in the estimation process—the so-called “treatment effect modifiers” (Athey & Imbens, 2016)—relies entirely on prior research concerning the primary drivers and determinants of household resilience (Upton et al., 2022). At the same time, CLARE's flexibility allows it to strike a balance in the trade-off between capturing a comprehensive representation of all relevant dimensions of resilience and identifying the minimum set of variables needed for robust out-of-sample prediction. This enables the targeting of interventions even in contexts where data collection is difficult, costly, or both, and priority choices must be made. Consequently, the construction of our indicator does not come at the expense of transparency and interpretability.

In this shock-focused modeling framework of resilience, shocks play a central role, and the estimation of CLARE entirely revolves around and incorporates shocks into the analysis from the outset, accounting for the protective role played by each analyzed resilience subcomponent in determining heterogeneous effects of the shock.³ As such, CLARE scores can be interpreted as “the capacity that ensures stressors and shocks do not have long-lasting adverse development consequences” (Constas et al., 2014). In this paper, we apply the methodology to build an indicator of *drought resilience*, using droughts as the key shock of interest and food security as the outcome. We also discuss and show how to combine multiple shocks and estimate a ‘general-purpose’ resilience indicator within the CLARE framework. More generally, CLARE can be adapted to any kind of shock affecting wellbeing and any wellbeing outcome.

³ It is important to recognize that if a significant portion of the harm caused by (uninsured) risk exposure arises from the behavioral adjustments households make to cope with the possibility of shocks, our measure (like other existing resilience indicators) may not fully capture these risks. In this light, the shock effects we estimate should be interpreted as a lower bound—a conservative approximation of the true impact of such shocks.

This versatility and explicit anchoring to shocks are novel compared to existing indicators and stem from the fact that, at its core, CLARE is grounded in an impact evaluation mindset, an aspect we share with other recent work on resilience (Alloush & Carter, 2024).

An important feature of CLARE is its scalability: once the weights and conditional probabilities have been estimated using causal ML, computing the index on new data becomes straightforward under the assumption of stability or invariance in the data-generating process underlying the relationship between shocks, outcomes, and resilience subcomponents. It suffices to predict the probability of falling above or below the normative threshold under the shock scenario out-of-sample, and to plug in the new values of the resilience subcomponents to calculate their weighted average using the pre-computed weights. The key point is that when computing the index on new data from another area (or even country)—and therefore aiming for cross-sectional prediction—the user must assume that the relative importance weights of the various components remain the same. The use of ML algorithms to predict on held-out test data can provide indirect evidence regarding the credibility of this assumption. For instance, a model trained on a set of countries and performing well on a testing set composed of observations from an unseen country would indicate the absence of distribution shifts, that the model generalizes well to new data, and can thus be deployed for transfer learning. Similarly, if the new data represent future data points (forecasting), the assumption is that the relationship between the wellbeing outcome, the shock, and the conditioning variables remains stable over time. The stability of the importance weights can also be assessed empirically, as we show in later sections.

Under these assumptions, CLARE can be employed for transfer learning in data-scarce environments where only cross-sectional data are available, or for forecasting purposes aimed at implementing crisis preparedness and early warning mechanisms in shock-prone areas for vulnerable households (or more aggregate units⁴). Additionally, estimating data-driven weights enables the establishment of an objective hierarchy among resilience components. This, in turn, offers agencies and policy makers the opportunity to prioritize, from a cost-effectiveness perspective, the most important targetable household characteristics that would yield the greatest increase in overall resilience capacity within a given context.

This work belongs to the quantitative strand of development resilience literature. Specifically, we draw on the methodological framework established by Upton et al. (2022)—and, although in a different setting

⁴ In principle, CLARE's methodology can be adapted to estimate resilience indicators for larger units of analysis, including communities and value chains.

with a focus on vulnerability, by Doan et al. (2023)—to construct and test our resilience indicator using nationally representative, multi-topic, *longitudinal* household survey data from the World Bank’s Living Standards Measurement Study (LSMS). These datasets were selected for their key attributes essential to constructing a resilience indicator, including their longitudinal structure, multi-topic coverage, data harmonization, and georeferenced data collection that allows integration with geospatial data on climate, conflicts, and environment that provide objective measures of shock exposure. Specifically, we analyze longitudinal survey data consisting of more than 35,000 initial observations and 19 survey waves from four Sub-Saharan African countries: Malawi, Nigeria, Tanzania, and Uganda—all of which currently prioritize resilience as a major policy topic. This dataset is considerably larger than those previously used in the specialized literature, offering a comparative advantage in terms of the external validity of our findings. Our results demonstrate that CLARE exhibits good out-of-sample performance, both in predicting food insecurity on data from held-out countries (cross-sectional prediction) and in forecasting food insecurity for observations in left-out future periods (dynamic forecasting). Across these tasks, CLARE consistently outperforms both the best-performing resilience indicator in the Upton et al. (2022) study—the empirical parametric specification derived from the Cissé and Barrett (2018) conditional moments-based approach—as well as other existing resilience indicators and alternative estimation approaches. CLARE’s better performance is likely ascribable to the ability of causal ML tools to flexibly capture household-level heterogeneity in the effects of covariate shocks on wellbeing due to differences in elementary resilience components, as well as their capacity to handle non-linearities and interactions in the underlying data-generating process (Hastie et al., 2009). Finally, we provide evidence of CLARE’s robust performance across multiple weather shock datasets and alternative sets of resilience components.

The main contribution of this paper is substantive: we introduce a new resilience indicator to the toolbox of resilience practitioners that is sufficiently flexible to be applied to different shocks and empirical contexts. As explained, CLARE is conceived as an explanatory variable intermediating between shocks and development outcomes and is estimated as a continuous indicator. However, CLARE can also yield a binary indicator to classify households as ‘resilient’ or ‘not resilient,’ which might be more practical for policy targeting purposes. It exhibits internal consistency and is, by design, indexed to a normative threshold or standard of wellbeing, aligning with another key definition of resilience: “the ability to achieve and maintain an acceptable standard of wellbeing even in the face of shocks and stressors” (Barrett & Conostas, 2014). Thus, CLARE can also serve as an outcome variable for targeting and

evaluation purposes.⁵ More broadly, we contribute to the development literature concerned with the aggregation of multidimensional wellbeing indices. These indices have often been criticized for the arbitrary choices involved in combining individual components into a composite indicator (Duclos et al., 2006). In this context, the idea of data-driven weights derived from ex-post evaluation, as outlined in this work, could offer a more objective approach to aggregating composite indicators beyond resilience.

The rest of the paper is arranged as follows. Section 2 presents the related literature and conceptual framework. Section 3 describes the methodology. Section 4 presents an empirical application of the method to build a composite indicator of drought resilience. Section 5 concludes.

2. Related literature and conceptual framework

This work is related to two main strands of literature: the methodological literature leveraging the use of ML algorithms in development, which we build upon and extend by proposing a causal ML approach to create a composite development indicator; and the theoretical and empirical literature on development resilience, from which we draw the theoretical foundations for the conceptual framework underlying our resilience indicator.

From a methodological viewpoint, we are the first to use causal ML algorithms to build an index of household resilience, and more generally, a composite indicator in the development field. Although ML has recently been extensively used in development to predict and map poverty and food security outcomes, also in conjunction with the use of non-conventional data sources (e.g., Aiken et al., 2022; Aiken et al., 2023; Baez et al., 2024; Barriga-Cabanillas et al., 2025; Blumenstock et al., 2015; Browne et al., 2021; Constenla-Villoslada et al., 2025; Echevin et al., 2024; Hossain et al., 2019; Jean et al., 2016; McBride & Nichols, 2018), causal ML tools have not been yet employed.⁶ In the resilience literature, only a handful of studies (Garbero & Letta, 2022; Knippenberg et al., 2019; Villacis et al., 2024) have employed ML algorithms in the context of resilience measurement. Specifically, Garbero and Letta (2022) employ a set of supervised ML algorithms to predict the subjective ability to recover from shocks using cross-country data from the International Fund for Agricultural Development. Knippenberg et al.

⁵ Using CLARE for subsequent impact evaluation means employing it as an outcome variable to estimate the impact of a given program on household resilience. At present, the only existing resilience indicator that can be used in this way is the resilience score of Cissé and Barrett (2018). The use of CLARE as an outcome variable for impact evaluation is, however, conditional on employing an appropriate approach for inference and accounting for uncertainty (e.g., bootstrapping), as the outcome would be an estimated rather than observed variable.

⁶ A recent exception is the study by Letta et al. (2024), who investigate heterogeneity in climate migration responses using causal forests.

(2019) use LASSO and random forest to identify the best predictors of the Coping Strategy Index (CSI), using high-frequency (monthly) data from Malawi. Villacis et al. (2024) proxy resilience capacity with the Food Insecurity Experience Scale and, using, two waves of phone survey data from the World Bank for four African countries, show that ML models can effectively predict resilience in a classification framework.

It is important to understand the fundamental difference between the use of ML in our study versus those that are discussed above. In the latter set, ML is employed to simply *predict* a given measure of resilience and uncover systematic patterns and ‘unusual variables’ among the predictors associated with higher or lower values of the outcome. These predictive models can only uncover associations and correlational patterns between the predictors of resilience and resilience capacity. However, in resilience analysis, causality matters: the relationship between shocks, wellbeing, and resilience drivers is causal, not merely correlational, and therefore demands rigorous causal inference techniques. Our *causal* ML approach is different because we use recent techniques that predict treatment effects, not outcomes, and are fully embedded in the potential outcomes framework of causal inference econometrics. In doing so, we estimate granular treatment effects for distinct subgroups defined by a set of conditioning variables that intermediate the relationship between treatment and outcome—in our case, the resilience subcomponents. The estimated CATEs are, in turn, indicative of underlying mechanisms (Knaus, 2022).

Therefore, causal ML is an ideal tool to estimate resilience because it unpacks the causal relationship between wellbeing and shocks and identifies capacities and components of resilience that insulate from or expose distinct subpopulations to the adverse consequences of shocks. Lastly, being tailored for longitudinal survey data, CLARE can track resilience dynamics more effectively than in Garbero and Letta (2022) and Villacis et al. (2024). At the same time, it does not require high-frequency data, whereas the predictive ML method proposed by Knippenberg et al. (2019) depends on it. This is important in terms of the scalability of the proposed indicator: high-frequency data are rare in development, and most survey waves typically occur every two to three years at best.

From a conceptual perspective, our proposed method aligns with multiple theoretical foundations. Specifically, CLARE addresses the two main typologies of resilience identified by Barrett et al. (2021), namely resilience as *ex-ante capacity* and resilience as a *normative condition*.⁷ The ex-ante capacity

⁷ Barrett et al. (2021) introduce a third notion of resilience: resilience as a return to equilibrium. This approach describes a condition—specifically, ex-post recovery from shocks—of a wellbeing variable of interest, rather than attempting to explicitly model the various capacities that ultimately lead to recovery. However, this conceptualization of resilience aligns more closely

conceptualization (i.e., the most commonly adopted approach) considers resilience as a latent variable that captures the combined effects of both observable and unobservable attributes of individuals, households, or larger units capable of mitigating the adverse impacts on wellbeing from both ex-ante risk and/or ex-post shocks (Béné et al., 2014). As is common for latent variables, measuring this typology of resilience remains an open challenge. In this context, some scholars (Quandt et al., 2019, Woolf et al., 2016) emphasize household resilience in terms of assets and argue that households scoring highly on a set of standard asset categories (including human and social ones) are more likely to be resilient. Others, such as FAO RIMA (FAO, 2016; D’Errico & Di Giuseppe, 2018; D’Errico et al., 2019) and TANGO (Smith and Frankenberger, 2018), propose the use of factor analysis to derive some operationalizable “resilience capacities” or “pillars” to serve as explanatory variables and present correlations with wellbeing outcomes. In the same vein, Béné (2013) further highlights the role of adaptive strategies in risk exposure.

As for resilience being conceptualized as a normative condition, Barrett and Costas (2014) and Cissé and Barrett (2018) define it as the individual probability of achieving some minimal standard of living conditional on a wide range of observable characteristics and shock exposure. The traditional limitations of this empirical measure of normative resilience are: (i) sensitivity to the choice of the adopted wellbeing measure; (ii) challenges in the empirical identification of shock exposure; and (iii) the confounding influence of unobservable factors, such as social capital, perceptions, and, more broadly, the role of qualitative and psychosocial factors (Carr, 2019). The probabilistic moments-based method by Cissé and Barrett (2018) provides an empirical framework to estimate resilience as a normative condition by integrating stochastic elements that capture the dynamics of resilience, making it particularly suited for policy and program evaluations, and has been recently applied and extended in several resilience measurement studies (Upton et al., 2016; Knippenberg et al., 2019; Upton et al., 2022; Scognamillo et al., 2023) as well as for impact evaluation (e.g., Premand & Stoeffler, 2022; Ranucci et al., 2025).

Conducting a comparative assessment of three mainstream resilience indicators using the longitudinal LSMS survey data for Ethiopia (across three survey waves from 2011 to 2016) and Niger (two waves between 2011 and 2015), Upton et al. (2022) argue that these indicators lack empirical validation. Specifically, they identified significant discrepancies in the distributions of the three indicators and their

with the concept of resilience as used in ecology and has limited utility in development practice, where scholars and practitioners typically address undesirable initial states such as poverty or food insecurity (Montalbano and Romano, 2022).

relative rankings in terms of household resilience. Moreover, they showed that these methods share disappointing out-of-sample performance. The best-performing indicator, which is by far the one developed by Cissé and Barrett (2018), achieved only modest overall accuracy in classifying food-insecure households in their application, whereas RIMA and TANGO performed no better than random chance. Furthermore, none of the resilience measures predicted significantly better than the far simpler approach of using the most recent observed value of wellbeing to predict future wellbeing. Therefore, Upton et al. (2022) argue, the value added in terms of predictive and targeting performance by these measures remains unclear. Overall, the authors conclude that “the approaches presently in play are all, at best, imperfect, and at worst deeply flawed.” It then comes as no surprise that, to date, resilience measures have been used largely to motivate development interventions, but much less for targeting or impact evaluation (Barrett et al., 2021).

Despite this disappointing evidence, the measurement of development resilience is at the forefront of a new wave of empirical research. This is driven by the growing demand for effective targeting methods and robust empirical evaluations to support resilience-building interventions. As part of this effort, Alloush and Carter (2024) propose a counterfactual approach to resilience measurement, which is rooted theoretically and empirically in the impact evaluation tradition and based on a quantitatively estimable resilience metric to relate to a no-shock counterfactual measure of household wellbeing. The introduction of this shock-free (and potentially risk-free) counterfactual benchmark allows them to address the limitations of Cissé and Barrett’s (2018) conditional moments-based approach, namely the reliance on a pre-defined level of wellbeing (e.g., the poverty line) which restricts a comprehensive analysis of resilience dynamics among non-poor households, as well as the sensitivity to the choice of the adopted wellbeing measure. Our work starts from a similar premise, in that the derivation of our data-driven weights is also based on the identification and estimation of household-specific counterfactual outcomes under the no-shock scenario and the associated conditional average treatment effects.

The limitation of using a predefined wellbeing measure has also been addressed by Lee et al. (2025). Recognizing that resilience measures constructed with different indicators yield inconsistent household orderings, which undermines the effectiveness of interventions aimed at improving development resilience, they explore how these outcomes can be aggregated through different weighting schemes to create a more comprehensive resilience measure rooted in the integration of (a) the probabilistic moment-based resilience measurement approach of Cissé and Barrett (2018) with (b) the multidimensional poverty measurement framework of Alkire and Foster (2011). This integration has the merit of enabling

a broader identification of development resilience. At the same time, the fact that univariate and multidimensional resilience measures exhibit different distributions and imply different household rankings leaves unresolved the critical issues related to project/policy targeting and impact evaluation. Their use of aggregation weights for different wellbeing proxies parallels our use of data-driven ML techniques to derive aggregation weights for the resilience components. Our work, with its focus on the aggregation of the explanatory components of the resilience indicators rather than outcome measures, is thus complementary to their approach and could potentially be integrated with it.

Against this background, CLARE advances the state of the art in this literature by using data-driven techniques to identify and evaluate the relative weights of the most relevant transmission channels that mediate the relationships between shocks, outcomes, and resilience subcomponents. This approach is fully transparent, flexible, and replicable. It enables the non-parametric empirical identification of nonlinear multiple equilibrium economic trajectories, fully aligning with the resilience and poverty trap literature and enabling the prioritization of alternative policy actions. In this respect, CLARE serves a dual purpose: it acts as a normative tool to classify households as resilient or non-resilient, and as an explanatory (or mediating) variable that synthetically connects shocks to development outcomes. In line with the most relevant empirical literature, it aims to balance the added informational value of incorporating new dimensions with the potential coverage loss that could impair out-of-sample performance. While the selected proxies may be context-specific and invite further debate on the most suitable indicators for robust resilience assessments, CLARE identifies the minimum set of proxies needed to support reliable out-of-sample evaluations—even in data-scarce environments.

3. Methodology

The construction of CLARE involves three sequential steps: (1) estimation, (2) aggregation, and (3) evaluation. Each of these steps is described in detail below.

3.1 Estimation

Constructing CLARE involves estimating two primary elements: (1) the data-driven importance weights of resilience subcomponents in driving the relationship between the outcome and the shock, and (2) the household-specific conditional probabilities of falling above or below the normative wellbeing threshold if the shock occurs.

To estimate these effects, we employ causal ML techniques. While various methods are suitable for this purpose,⁸ we focus on *causal forests*, which are designed to estimate Conditional Average Treatment Effects (CATEs) (i.e., non-parametric estimates of treatment effect heterogeneity based on selected covariates) (Athey & Imbens, 2016; Athey et al., 2019; Wager & Athey, 2018). They adapt random forests—a powerful supervised ML method used to predict outcomes from a set of inputs (Breiman, 2001; Hastie et al., 2009)—to the task of predicting heterogeneity in causal effects. Similarly, for the variable weights, we rely on the importance ranking produced by this specific method. However, several alternative importance measures are available in the interpretable machine learning field. Most of them are model-agnostic (e.g., SHAP values, permutation feature importance) and can also be used to aggregate CLARE scores.⁹

The causal forest algorithm creates an ensemble of causal trees. Each causal tree is defined by data-driven sample splits that form leaves, followed by predictions of causal effects in the terminal nodes based on a set of conditioning characteristics, the so-called ‘treatment effect modifiers’ (Athey & Imbens, 2016). The goal is to split the data to maximize effect heterogeneity across leaves. To prevent the risk of overfitting, the algorithm uses the “honest approach” (Athey & Imbens, 2016), which randomly splits the sample into two: one half (the prediction sample) defines the sample splits, and the other half (the estimation sample) estimates the predicted CATE. This is repeated for as many trees as there are in the forest (2,000 in the case of our application). Each tree identifies subgroups where treatment effects differ most based on candidate covariates, and the final prediction is a weighted average of predictions across trees. This average is consistent and asymptotically normal (Wager & Athey, 2018). Furthermore, causal forests can be applied to longitudinal data from observational studies—as in our case—or randomized control trials – for the estimation of heterogeneous treatment effects (Athey et al., 2023; Britto et al., 2022; Knittel & Stolper, 2025; Letta et al., 2024; Miller, 2020; Johnson et al., 2023; Zhang & Luo, 2023).

Before building our model, we split the sample into two distinct sets: a training set, where we estimate resilience components’ weights and develop our model, and a testing set, where we evaluate the model’s predictive performance. The definition of these sets depends on the prediction goal. If the goal is forecasting, the training set should include earlier waves, with the testing set comprising later ones. For cross-sectional predictions, such as the ‘out-of-country’ predictions, the training set should rotate to include all-but-one countries, with the testing set being the left-out country.

⁸ For a comprehensive review of such methods, see Chernozhukov et al. (2024).

⁹ See Molnar (2020) for an overview of these techniques.

In our longitudinal setting, we model the relationship between the outcome, shock, resilience components, and confounders as shown in Equation 1:

$$W_{it} = \tau(\mathbf{Z}_{it-1})S_{it} + f(\mathbf{X}_{it-1}, \bar{\mathbf{X}}_i, \bar{\mathbf{X}}_{.t-1}) + \varepsilon_{it} \quad (1)$$

where W_{it} represents a binary wellbeing outcome anchored to a normative threshold¹⁰—i.e., in our illustrative application, ‘Food insecurity’ status, i.e., defined as a Food Consumption Score equal to or below 35, using the same threshold as Upton et al. (2022)—for household i in survey wave t ; S_{it} is the binary variable capturing exposure to shocks, defined by either a single type of shock (e.g., drought, flood, conflict, etc.) or the occurrence of *at least one* (or another threshold) among multiple shocks. \mathbf{X}_{it-1} is a vector of time-varying confounders measured at time $t - 1$, which should also include some or all of the resilience subcomponents if the researcher believes that they can autonomously influence the outcome, independent of and in addition to their protective role in relation to shocks; $\bar{\mathbf{X}}_i$ and $\bar{\mathbf{X}}_{.t-1}$ are Mundlak-type controls including unit-specific and wave-specific averages of all time-varying controls capturing time-invariant and time-varying unobserved heterogeneity¹¹; and $\tau(\mathbf{Z}_{it-1})$ captures the heterogeneous effects of the shock, which depend on the household-specific values of a number (which we later index with j) of resilience subcomponents \mathbf{Z} measured at time $t - 1$, ensuring that they are not affected by the shock and that there is no ‘bad control’ issue at play.

This partially linear model aims to non-parametrically estimate the shock effect $\tau(\mathbf{Z}_{it-1})$ conditional on \mathbf{Z}_{it-1} , with $f(\mathbf{X}_{it-1}, \bar{\mathbf{X}}_i, \bar{\mathbf{X}}_{.t-1})$ capturing confounding effects. It strikes a balance between structure and flexibility, as the effect of the shock is additive and interpretable, while the data-generating process for the untreated potential outcome can exhibit almost arbitrary complexity (Chernozhukov et al., 2024)—a complexity we address using fully non-parametric methods. The shock variable is assumed to be independent of potential outcomes conditional on controls (Athey et al. 2019). The main identifying

¹⁰ The use of a binary outcome in the estimation is necessary to anchor the outcome to a normative threshold, as recommended by Upton et al. (2022), and to allow for the estimation of conditional probabilities of falling below (or above) such a normative threshold. However, CLARE is later estimated as a continuous indicator, which can be binarized and compared with either a continuous or binary wellbeing outcome, as we will show in the application.

¹¹ To capture unobserved heterogeneity (a.k.a. fixed effects in a parametric setting), we include Mundlak-type controls (Mundlak, 1978). This approach involves incorporating unit- and time-specific averages of time-varying covariates as a common solution for controlling both time-invariant and time-varying unobserved heterogeneity—not only in standard predictive ML but also in causal ML methods and other new causal inference techniques designed to account for group-level unobserved heterogeneity (Arkhangelsky & Imbens, 2024; Chernozhukov et al., 2024; Wooldridge, 2021). To operationalize this approach, we use the method for sufficient representation of categorical variables proposed by Johannemann et al. (2019), which is implemented via *sufrep*, a sister package of *grf*.

assumption underlying the causal forest approach is thus conditional unconfoundedness. To ensure its credibility, we follow Wager and Athey (2018), who recommend a double orthogonalization approach. Specifically, we regress out the main effects of \mathbf{X}_{it-1} , $\bar{\mathbf{X}}_i$, $\bar{\mathbf{X}}_{t-1}$ on W_{it} and S_{it} and retain the residuals, W_{it}^* and S_{it}^* , to build causal forests.

This step involves estimating two preliminary and fully non-parametric regression forests on the training set, where we can include any high-dimensional set of covariates that may confound the causal relationship between the outcome, the shock, and the resilience components. In doing so, we ensure that the residualized outcome and shock variables are filtered from time-invariant household characteristics, time-varying lagged household characteristics, and time trends.

This approach provides multiple layers of robustness to ensure the plausibility of the unconfoundedness assumption. Consider the case of our application on resilience to drought. First, weather shocks are generally assumed to be random. Second, the SPEI is normalized with respect to normal local conditions, thus filtering out heterogeneity due to different local averages. Furthermore, the double orthogonalization approach we adopt (prior to impact estimation) non-parametrically controls for confounders and eliminates any residual sources of confounding effects and omitted variable bias. Finally, to capture heteroskedasticity in ε_{it} , one should rely on cluster-robust estimation, such as clustering at the household level, at all estimation steps. Cross-validation of all tuning hyperparameters of the forests should also be conducted.

After applying orthogonalization and residualizing the outcome and shock variables, we proceed with estimating the CATEs, $\tau(\mathbf{Z}_{it-1})$. The estimation of heterogeneous effects involves a residual-on-residual non-parametric regression using W_{gt}^* and S_{it}^* , obtained from the previous double orthogonalization step, conditioning on a set of J resilience components \mathbf{Z}_{it-1} that determine heterogeneity in the shock effects. Since our outcome variable is binary, the estimated effects must be interpreted as the change in probability of the outcome occurring that is associated with the shock (in our case, the drought occurrence). Combined with the orthogonalization of the outcome, the estimation of the household-specific effects also allows us to estimate the unit-level potential outcome under the shock scenario, i.e., the conditional probability of falling below the normative threshold if the shock occurs.

These household-specific or group-level effects can be mapped and analyzed to assess the heterogeneity at play and the role of the main resilience determinants. While estimating causal effects is important, it

is not our primary focus. Instead, we concentrate on two key elements needed to estimate CLARE: (i) the probability of being above the normative threshold if the shock occurs, which is the complement of the probability of being below the threshold (e.g., food insecure) under the shock; and (ii) the variable importance ranking of the causal forest, based exclusively on training set data, which provides estimated data-driven importance values for each resilience subcomponent in mediating the causal relationship between the shock and the outcome. The importance weights for each resilience subcomponent are derived from the variable importance measures generated by the causal forest model. The ‘importance’ of a variable is measured by the frequency with which a resilience component j is split on. Interpreting importance measures as “weights” means that they reflect how significantly each variable contributes to mitigating the negative effects of shocks on household wellbeing.¹² In the causal forest approach, the data is recursively partitioned based on conditioning characteristics to form subgroups with heterogeneous effects. At each split in the tree, the algorithm selects the variable that best separates the data into groups with differing treatment effects. This process is repeated across all trees in the forest, and the frequency with which each covariate is used in these splits indicates its relative importance. Formally, the importance weight for covariate j is the share of splits across all trees in the forest that select j as the splitting variable. This weight, w_j , reflects the role of the j -th variable in determining overall resilience of households. The higher the frequency of a resilience component appearing in the trees, the more crucial it is in explaining how different households respond to the same shock. We use convex relative importance weights, which are non-negative and sum to one.¹³

Crucially, the estimation of the data-driven weights is carried out exclusively using training set data, not on the testing data (later waves or other countries, respectively, for forecasting and cross-sectional prediction) on which CLARE’s performance will be tested. The entire procedure for developing the causal forest is run only on the subset of training data, so the feature importance ranking reflects the role played by each component in determining the impact of the shock on the outcome in the *training dataset*. This is important because it allows us to ‘export’ the weights to other data, plug in new values for the

¹² Some variables might actually amplify rather than mitigate the effects of the shock, for instance, distance from the market. In these cases, as we later explain, we use the estimated weight but take the complement of the standardized variable in the final aggregation, so that, as in the example, the complement of distance from the market contributes positively to the overall resilience score, with the weight estimated by the forest.

¹³ An important caveat pertains to the issue of correlated variables included in the set of treatment effect modifiers. Most ML techniques that establish an importance ranking among features—including the variable importance measures produced by the causal forest, which we focus on here—are sensitive to correlated features, in the sense that adding a correlated variable can reduce the importance of the associated feature by splitting the attribution between them (as discussed in Molnar, 2020). This may result in low importance for both features when included together, but high for each when the correlated one is excluded from the model.

resilience components of these new data, and compute their weighted average on held-out observations. The underlying assumption is that the weights remain consistent across different datasets and that the input-to-output relationships estimated in the training set also hold true in the new environment. For forecasting, this translates into temporal stability of the estimated relationships; for cross-sectional prediction on data from other countries, it means that once country fixed effects and many other confounders are filtered out, the weights of the different resilience components and their hierarchy are uniform across countries.

Lastly, while the weights are later used for variable aggregation, the prediction of the causal effects and the conditional probabilities of falling below the threshold—used to rescale the weighted average components (see the next subsection)—are carried out out-of-sample for the new observations at the end of the causal forest estimation, provided that the same set of variables employed in the training set is available for the testing set. In our case, we produce out-of-sample predictions for the full sample, including the testing set, and not merely for the latter, because we want to estimate CLARE for the entire sample. The out-of-sample performance of CLARE, however, will be tested exclusively on the testing data, which the model has never seen during the training phase.

Before moving to the final computation of CLARE, two additional remarks are in order. First, unlike other resilience estimation methods, the approach described above is fully flexible and non-parametric, without making any functional form assumptions regarding the underlying data-generating process. This flexibility allows for the smooth estimation of the non-linear welfare relationships that characterize wellbeing and resilience dynamics, as described in Section 2 and emphasized in the literature on non-linear wellbeing and poverty dynamics and asset-based poverty traps (Barrett et al., 2016; Carter & Barrett, 2006). In contrast, this non-linearity is only partially accounted for in other resilience indicators, such as the resilience score of Cissé and Barrett (2018), through the inclusion of wellbeing polynomials.

Second, this methodology is not necessarily tied to the use of a specific method, namely the causal forest. Other causal ML methods can be employed to estimate the quantities of interest, as long as they (i) estimate household-specific conditional average treatment effects and probabilities, and (ii) produce variable importance rankings or similar outputs (e.g., SHAP values (Lundberg & Lee, 2017) or Local Surrogate Models (Molnar, 2020) from the Explainable AI literature) that allow for the interpretation of machine predictions and can be used to derive data-driven weights for the resilience components.¹⁴

¹⁴ Refer to Chernozhukov et al. (2024) for a review of alternative causal ML methods suited for treatment effect heterogeneity.

3.2 Aggregation

In the second step of estimation, the importance weights and conditional probabilities are aggregated into the final CLARE indicator. Specifically, CLARE is computed as a weighted sum of pre-shock values of the J resilience subcomponents, which must be preliminarily normalized in the 0-1 range to be aggregated.¹⁵ Each subcomponent j of the vector of resilience components \mathbf{Z} for household i is multiplied by its corresponding importance weight w_j , derived from the causal forest. The weighted subcomponents are then aggregated by taking their sum and rescaled by the probability of being above the normative threshold (food secure) if the shock occurs. More specifically, the formula for the computation of CLARE is presented in Equation 2:

$$CLARE_{it} = \left(\sum_{j=1}^J \omega_j \mathbf{Z}_{it-1} \right) * \left(1 - Pr(W_{it} \leq \underline{W} \mid S_{it}, \mathbf{X}_{it-1}, \bar{\mathbf{X}}_i, \bar{\mathbf{X}}_{t-1}) \right) \quad (2)$$

where $\left(\sum_{j=1}^J \omega_j \mathbf{Z}_{it-1} \right)$ is the data-driven weighted average of the J lagged resilience components, and $\left(1 - Pr(W_{it} \leq \underline{W} \mid S_{it}, \mathbf{X}_{it-1}, \bar{\mathbf{X}}_i, \bar{\mathbf{X}}_{t-1}) \right)$ is the conditional probability of falling above the normative wellbeing threshold under the shock scenario. This aggregation approach ensures that subcomponents with greater relevance to resilience, as identified through the causal forest, are weighted more heavily in the final indicator. Rescaling by the conditional probability of being food secure under the shock is done because the weighted average of the components pertains to the change in wellbeing occurring due to the shock; it is therefore not necessarily correlated with wellbeing levels. However, resilience indicators should be correlated with levels other than changes in wellbeing (Upton et al., 2022). Since both the normalized weighted average and the probabilities fall in the 0-1 range, so does their product, CLARE. We then multiply it by 100 for interpretability, as is done with other resilience indicators.

Note that we index CLARE at time t , but both the elements on which the index is computed are created based on information available only a $t-1$ or (in the case of time trends) that does not require data collection and availability also in the current wave. This is because CLARE is designed to be forward-

¹⁵ Resilience components are measured on different scales. Therefore, to allow for comparability between components, the variables must be standardized before aggregation—typically by normalizing them to have a mean of 0 and a standard deviation of 1. Additionally, all components of a composite indicator must have the same polarity, i.e., the sign of the relationship between the component and the phenomenon of interest. Therefore, standardized variables that exhibit a negative correlation with the food security outcome (e.g., distance from markets or a dummy for living in rural areas) are taken as their complements so that they contribute positively to resilience capacity, while maintaining their absolute correlation value with the outcome. This is in line with established practice for aggregation in mainstream indicators such as RIMA and TANGO.

looking with respect to wellbeing status and applicable for anticipatory policy targeting. For example, a high CLARE score in the current wave indicates that, based on data from previous waves, a household is likely to remain above the normative threshold if exposed to a shock. A binary classification can be constructed using different criteria depending on the application and the preferred trade-off between false positives and false negatives, while also accounting for cost-effectiveness considerations. In the application, we follow Lee et al. (2025) and use the median CLARE score to differentiate between resilient and non-resilient households.

3.3 Evaluation

Rigorous out-of-sample testing of the predictive performance of any model must be carried out on previously unseen data, on which the model was not trained. This introduces a “firewall” principle: none of the data involved in generating the prediction function is used to evaluate it (Mullainathan & Spiess, 2017). The performance of the model on the held-out data of the testing set can be considered a reliable measure of the “true” performance (Hastie et al., 2009). We are interested in two out-of-sample tests:

- i) Forecasting wellbeing status for observations in left-out future periods using CLARE values computed based on quantities estimated only using data from previous waves.
- ii) Predicting the wellbeing status of households from held-out (testing) countries using CLARE values computed with weights estimated on household data from different training countries (cross-sectional prediction); we label this predictive task as ‘out-of-country’ testing.

Our sample splits are not random as in standard ML but are instead based on time or country to avoid “data leakage” (Cerqua et al., 2025) when applying ML to longitudinal data—that is, contamination of the training set that could lead to overly optimistic assessments. After estimating the quantities of interest separately for each task, CLARE’s out-of-sample performance can be evaluated using a range of metrics appropriate for both regression (e.g., Mean Squared Error (MSE), R-squared, correlation coefficients) and classification tasks (e.g., confusion matrices, AUC-ROC). Its performance can then be compared to that of other existing indicators.

The full pipeline for CLARE is summarized in Table 1. In the next section, we apply this pipeline to the longitudinal survey data described in Section 4, for both out-of-sample forecasting and out-of-country prediction of food insecurity.

Table 1: The CLARE pipeline

<p>1. Preliminary – Data splitting</p>
<p>We split the full dataset into a training set, on which we will compute resilience weights and conditional probabilities of falling below the normative threshold under the shock scenario, and a testing set, on which we will later test the out-of-sample performance. The definition of the training and testing sets depends on the prediction goal: if the goal is forecasting, the training set should include earlier waves, and the testing set should include only later ones; if the goal is cross-sectional prediction—e.g., out-of-country—the training set should be a rotating one that includes all-but-one countries, and the testing set should be the left-out country.</p>
<p>2. Estimation – <u>Use only the training set</u></p>
<p>In this step, we use causal ML tools (e.g., causal forests) to estimate the heterogeneous effects of the shock (e.g., droughts) on a binary wellbeing outcome (e.g., food insecurity status), conditional on the pre-shock values of the resilience subcomponents. Non-parametric estimation based on double orthogonalization removes confounding effects, while the estimation of heterogeneous effects is based on a residual-on-residual regression conditional on the resilience subcomponents. At the end of this step, we derive the data-driven weights from the variable importance ranking for the resilience subcomponents.</p>
<p>3. Prediction</p>
<p>Armed with the model estimated on the training set, we predict the outcome on the full sample, which includes testing set observations. This allows to derive conditional probabilities of falling above the normative threshold under the shock, as well as the heterogeneous, household-specific effects of the shock.</p>
<p>4. Aggregation</p>
<p>We then compute CLARE as follows: i) Weight aggregation: We take a household-specific weighted average of the normalized resilience subcomponents, with weights derived from step 2; ii) Rescaling: We rescale this weighted average by the probability of falling above the threshold under the shock, as estimated in step 3; iii) Index scaling: We multiply by 100 to obtain a 0-100 index; iv) Discretization: the continuous indicator can be discretized into a binary one using different criteria. This process is conducted for the full sample to estimate the resilience indicator for all observations; however, out-of-sample performance testing must be carried out exclusively on the testing set data points.</p>
<p>5. Evaluation – <u>Use only the testing set</u></p>
<p>We assess the out-of-sample performance of CLARE in predicting the wellbeing outcome on the left-out testing set (later waves in forecasting; countries in out-of-country prediction) using a variety of common performance metrics for both regression and classification tasks using, respectively, the continuous and binary pairs of outcome and indicator. We also compare CLARE’s performance with that of other mainstream measures of resilience and alternative approaches.</p>

4. Empirical application

4.1 Data

4.1.1 Survey data

The survey data for our analysis come from the World Bank Living Standard Measurement Study (LSMS)-supported longitudinal surveys conducted by the national statistical offices in Malawi, Nigeria, Tanzania, and Uganda over 2009-20.¹⁶ There are several key features of these surveys that make them essential for constructing a resilience indicator. First, longitudinal data are essential because resilience is inherently dynamic and tracking households over time enables the identification of causal relationships between shocks, resilience capacities, and wellbeing. Second, multi-topic surveys capture the multi-dimensional nature of resilience, integrating data on household demographics, assets, energy access, education, livelihoods and community-level factors—all of which are critical for comprehensive measurement of resilience. The detailed data collection on household agricultural activities is also crucial for resilience analysis, particularly in developing contexts where agriculture plays a central role as the main transmission channel of the impact of climate shocks on welfare outcomes. Third, the surveys rely on cross-country comparable data collection methods, facilitating the creation of harmonized datasets that in turn enhance the external validity of the study and allows testing out-of-sample performances not only on future waves of the same country (forecasting), but also on left-out countries in a cross-sectional prediction test. The latter is important to ensure confidence in applying CLARE to new data from countries beyond those used in its development. Fourth, the surveys georeference household locations, which allow integration with geospatial data sources, such as remotely sensed climate data that provide objective measures of exposure to extreme weather events.

The choice of countries reflects computational and technical factors, such as balancing broad geographic coverage to ensure external validity while meeting the minimum data requirements—such as the number of survey waves—necessary to apply ML algorithms and develop the resilience indicator. The variables across the different countries are consistently constructed based on the available information, ensuring data comparability, which serves two purposes. First, it ensures that, by using a single unified dataset for all countries, we maintain consistency in the information, which is vital for accurate cross-country

¹⁶ The surveys had been conducted under the LSMS-Integrated Surveys on Agriculture (LSMS-ISA) initiative (<https://www.worldbank.org/en/programs/lsmis/initiatives/lsmis-isa>) and are publicly available from the World Bank Microdata Library.

comparisons. Second, the external validity of the indicator’s performance is higher compared to previous studies that relied on more limited datasets. The final dataset reflects a balance between maximizing the inclusion of resilience components across all countries and minimizing sample size losses. While each country provides the same core information, several differences exist in the number, composition, and timing of survey waves. Table 2 provides an overview of the survey data, including the number of waves corresponding to each country.¹⁷ The dataset construction started with 35,000 initial observations, including data from the baseline survey of each country. After creating variable lags for the second survey wave, the final analysis sample includes 28,112 households.

Table 2: Sources of Household Data

Country	Survey name	Years	Final n
Malawi	Integrated Household Panel Survey (IHPS)	2010/2011	<i>baseline</i>
		2013	1,282
		2016-2017	1,823
		2019	1,719
Nigeria	General Household Survey (GHS)	2010/2011	<i>baseline</i>
		2012/2013	3,619
		2015/2016	3,650
		2018/2019	1,219
Tanzania	Tanzania National Panel Survey (TZNPS)	2010/2011	<i>baseline</i>
		2012/2013	317
		2014/2015	687
		2019/2020	654
Uganda	Uganda National Panel Survey (UNPS)	2009/2010	<i>baseline</i>
		2010/2011	2,214
		2011/2012	2,211
		2013/2014	977
		2015/2016	2,607
		2018/2019	2,666
		2019/2020	2,467
Total	4 countries	19 waves	28,112

¹⁷ In **Malawi**, we use the *Integrated Household Panel Survey 2010-2013-2016-2019 (Long-Term Panel, 102 EAs)* implemented by the Malawi National Statistical Office with support from the LSMS-ISA initiative. For **Nigeria**, we employ the *Uniform Panel Dataset* which is derived from the past rounds of Nigeria’s General Household Survey – Panel (GHS-Panel) implemented by the Nigeria National Bureau of Statistics with support from the LSMS-ISA initiative. The panel covers the waves 2010-2011, 2012-2013, 2015-2016, and 2018-2019. For **Tanzania**, we created a panel dataset using the waves coming from the Tanzania National Panel Survey (TZNPS) implemented by the Tanzania National Bureau of Statistics (NBS) with support from the LSMS-ISA initiative in 2008-2009, 2010-2011, 2012-2013, 2014-2015, and 2019-2020. We exclude the first wave from our analysis since it does not contain information to construct our main outcome variable namely the Food Consumption Score (FCS). For Tanzania, a large number of observations from the original sample were lost due to extensive missing data across key resilience components, leading to a small final sample size. For **Uganda**, the data come from the Uganda National Panel Survey (UNPS) implemented by the Uganda Bureau of Statistics with support from the LSMS-ISA initiative. We constructed a panel using the rounds of data collected in 2009-2010, 2010-2011, 2011-2012, 2013-2014, 2015-2016, 2018-2019, and 2019-2020. For all countries, we retained households observed in at least three time periods—one required to compute lagged resilience components, and the others for minimum estimation in a panel context—for inclusion in the final dataset.

4.1.2 Outcome variable

CLARE is intended to be explicitly indexed to a measure of wellbeing. In this study, we use the Food Consumption Score (FCS) (Wiesmann et al., 2009) as a key wellbeing metric. The reason for this choice is twofold: first, this is the benchmark indicator of food insecurity employed in the comparative study by Upton et al. (2022); second, because it is the only measure consistently available across all waves and countries (different from the food consumption variable) ensuring comparability over time and across countries. The FCS is a frequency-weighted measure of dietary diversity, commonly referred to as a "food frequency indicator". It is calculated by analyzing how often a household consumed items from eight specific food groups during the seven days preceding the survey. The process of constructing the FCS begins by grouping food items into designated categories. For Malawi, Nigeria, and Tanzania, this grouping was consistently applied using a module provided by the respective data sources. However, in the case of Uganda, since the food categories were not predefined, we manually grouped the individual food items before calculating the FCS. The frequency of consumption for each food group is then summed, with a maximum limit of seven instances per group. Each food group score is weighted according to its nutritional importance,¹⁸ and the weighted scores are summed to produce the overall FCS. Following Upton et al. (2022), this continuous score is then converted into a categorical variable capturing food insecurity status: households with $FCS \leq 35$ are considered food insecure. Finally, note that this outcome, being related to food consumption, is harder to predict compared to household wealth (Barriga-Cabanillas et al., 2025).

4.1.3 Resilience components and confounders

As noted earlier, this analysis is conducted using a unified dataset, ensuring consistency across all countries by constructing an identical set of variables. To maintain methodological rigor and comparability, we closely align our variable selection with the framework utilized by Upton et al. (2022), which builds upon the methodological approach developed by Cissé and Barrett (2018). The variables were carefully matched to reflect data availability within the LSMS surveys for Malawi, Nigeria, Tanzania, and Uganda, ensuring that the dataset aligns with theoretical and empirical standards.

Specifically, the set of raw resilience components used as explanatory variables for this analysis includes household characteristics, such as the age, sex, and education level of the household head, as well as the

¹⁸ The eight groups, with the corresponding weight in parenthesis, are the following: main staples (2), pulses (3), vegetables (1), meat and fish (4), fruits (1), milk and dairy (4), sugar (0.5), oil (0.5).

proportion of household members with education levels exceeding primary and secondary schooling. Furthermore, we include measures of household participation in various livelihood activities, along with household size and composition—disaggregated by age and sex groupings to accurately capture the demographic structure—while also accounting for whether the household is in a rural or urban area. To assess household wealth, we include Tropical Livestock Units (TLUs) as a measure of livestock holdings, alongside an asset index derived through Principal Component Analysis (PCA), which is standard practice in recent development literature to capture long-term deprivation (e.g., Aiken et al., 2023; Ratledge et al., 2022). In addition, we incorporate community-level indicators, including access to basic services and the distance to the nearest market. In total, we include 18 raw resilience subcomponents, and this identical set is used to maintain comparability—although in different forms, such as lags versus contemporaneous values—for the construction of both CLARE and the resilience score of Cissé and Barrett (2018), which we employ for the comparative assessment.¹⁹

The description and summary statistics of the dependent variable FCS and the resilience components, as presented in Tables A.1 and A.2 in Appendix A, are based on the full sample including all observations for which FCS and its first lag are not missing. However, starting from this full sample, method-specific datasets are subsequently derived according to the specific requirements of each approach. This explains the differences in sample sizes across the tables presented in the following sections, as each method imposes distinct criteria for inclusion. Due to varying missing data patterns and/or variable selection for different approaches, the resulting analyses are conducted on datasets with heterogeneous sample sizes. For instance, CLARE construction relies exclusively on the use of lagged resilience components, whereas the Cissé and Barrett (2018) method uses mostly contemporaneous resilience variables. Consequently, there are differences in missing data between lagged and contemporaneous values of the same variables. The set of raw resilience subcomponents that will be employed to construct the different resilience indicators, however, is exactly the same.

Regarding the set of confounders included in the double orthogonalization of the outcome and shock variable before proceeding with the causal forest estimation, we include the pre-shock values of all 18 resilience components, as these variables also affect wellbeing independently of their intermediating role

¹⁹ Concerning feature correlation among the resilience subcomponents (cf. Footnote 13), only two variables—number of household members with primary education and number of household members with secondary education or higher—exhibit a strong correlation above 0.7 (specifically, 0.78), while most other pairwise combinations show minimal correlation. We retain both variables in the set of components to remain consistent with Upton et al. (2022), though readers should note that this may lead to an underestimation of the overall importance of education (Molnar, 2020).

with respect to the shock. This means that, for this application, the vector \mathbf{Z}_{it-1} includes the *same* variables as the vector \mathbf{X}_{it-1} . The reason is that these variables matter not only for resilience capacity—by intermediating the relationship between the shock and the outcome—but also because they can directly affect food security, independently of whether a shock occurs. Additionally, we account for time-invariant and time-varying unobserved heterogeneity by including Mundlak-type controls consisting of household-level and wave-specific averages of all these variables (Chernozhukov et al., 2024; Wooldridge, 2021). In our final sample, we keep all observations with complete outcome and covariate data.

4.1.4 Shock data

Since shocks play a central role in the CLARE framework, in this illustrative application we experiment with different shock definitions and measures. As noted above, CLARE can be computed as either a shock-specific or general-purpose resilience index. Our focus here is on constructing a composite indicator of *drought* resilience; We also report the performance of CLARE scores based on exposure to several different shocks.

To incorporate drought shocks, we integrate georeferenced household survey data with remote sensing weather data obtained from publicly-available third-party sources. As a drought indicator, we use high-resolution data on the Evaporative Demand Drought Index (EDDI) (Hobbins et al., 2016), a recent, physically based, multi-scalar drought index that complements precipitation-driven indicators by focusing exclusively on anomalies in atmospheric evaporative demand (E_0). E_0 is computed using the standardized reference evapotranspiration (ET) equation developed by the American Society of Civil Engineers (ASCE) (Walter et al., 2000), which is grounded in the physically based Penman–Monteith formulation (Monteith, 1965). This approach combines radiative and aerodynamic components to estimate atmospheric evaporative demand, relying on meteorological inputs such as temperature, humidity, wind speed, and solar radiation. It provides a consistent and widely accepted method for calculating reference ET under standardized surface conditions, typically assuming a well-watered reference crop and offering spatial resolution of approximately 0.125° (~ 12 km). EDDI values are standardized using a non-parametric inverse normal transformation, resulting in values that typically range between -2.09 and $+2.09$ (based on a 36-year climatology). Positive EDDI values indicate drier-than-normal atmospheric demand and signal the potential onset or intensification of drought, with values exceeding $+1$ and $+2$ corresponding to moderate and extreme evaporative anomalies, respectively. Due

to its sensitivity to short-term changes, EDDI is particularly effective in capturing both flash and sustained droughts, often offering earlier warnings than traditional supply-side indices. To integrate EDDI data with our main household dataset, we average EDDI values at the survey enumeration area (EA)-level and match them to households using GPS coordinates.²⁰ Notably, we use the true, confidential household geocoordinates rather than publicly available, spatially anonymized (obfuscated) GPS coordinates. This distinction matters because recent research shows that high-resolution remote-sensing weather products can be sensitive to spatial anonymization, which may introduce measurement error (Michler et al., 2022). We define a drought month as one in which the EDDI exceeds +1, indicating moderate drought conditions.

To capture weather conditions that matter for agricultural outcomes and, in turn, wellbeing changes, we focus only on *growing-season* EDDI (GS-EDDI), defined as the EDDI values during the growing season months in the period between two waves. To identify growing periods, we rely on crop calendars—differentiated by region—provided by the Famine Early Warning System Network (FEWS NET) of the United States Agency for Agricultural Development (USAID). Since, in each wave and visit, households are interviewed in different calendar months, we focus on household-specific GS-EDDI values depending on each household's interview date. This means that, for each household, we examine the weather conditions where they reside during all growing season months between their last and current interviews. We then construct a binary shock indicator from this continuous weather data, capturing drought exposure. We build this variable as follows: our shock dummy takes value 1 if the household experienced a drought (i.e., GS-EDDI above +1) during a share of growing season months in-between two waves exceeding the 75th percentile of the distribution, and 0 otherwise. This choice aligns with previous literature (Harari & La Ferrara, 2018) and is motivated by the consideration that a simple average of GS-EDDI values across all growing season months would obscure significant variation in weather conditions relevant to household wellbeing. In the main sample, the 75th percentile of the share of growing season months with drought occurrence is 12 percent. Accordingly, households are classified as exposed to shocks if they experienced droughts during more than 12 percent of the growing season period between two survey waves (cf. Table A.2 in Appendix A). This constitutes our main shock variable, whose impacts on FCS, as driven by the resilience subcomponents, will be estimated using the methodological pipeline described in Table 1.

²⁰ Values are extracted at the Enumeration Area (EA) level for non-mover households and at the household level for mover ones.

Recent research by Josephson et al. (2025) shows that estimating the relationship between weather conditions and socio-economic outcomes, specifically agricultural productivity, based on longitudinal survey data, is extremely sensitive to the source of weather data employed and emphasizes the importance of demonstrating that key results do not depend on the chosen weather data product. Therefore, for robustness, we also employ the Standardized Precipitation Evapotranspiration Index (SPEI), a multi-scalar drought index developed by Beguería et al. (2014) and publicly available in the Global SPEI database. The SPEI jointly considers precipitation, potential evaporation, and temperature, which provides a distinctive advantage over simple rainfall or temperature indicators as it incorporates the interaction between these variables in determining farmers' agricultural outcomes (Bertoli et al., 2022; Harari & La Ferrara, 2018). It is increasingly used in the literature to capture agriculture-relevant weather shocks and has been found to outperform other indicators in predicting crop yields (Vicente-Serrano et al., 2012). The SPEI is obtained by taking the difference between precipitation (P) and potential evapotranspiration (PET): $D_i = P_i - PET_i$. Then, D_i is standardized so that the indicator represents the deviation from the normal water balance (a SPEI of 0 indicates a value corresponding to 50% of the cumulative probability of D , according to a log-logistic distribution). A negative SPEI value is associated with dry events (lower rainfall) while positive SPEI values capture wet events (higher rainfall). The resolution of these SPEI data is lower (0.5*0.5 degrees) and, to construct the shock exposure variable, we adopt the same procedure described above for the EDDI data.

Finally, we illustrate how to construct a resilience indicator based on multiple shocks. In the recent development resilience theoretical framework (Barrett & Constanas, 2014), households and individuals in high-risk settings are not necessarily concerned with the specific source of the risk to which they are exposed, as many risks are correlated across both time and space. It is therefore important to allow for the construction of resilience scores based on multiple exposures that better reflect real-life vulnerability to a variety of shocks and stressors. To this end, we replace the drought shock variable defined above with a shock dummy equal to 1 if, during the period between two survey waves, the household experienced at least one of the following shocks:

- Drought shocks (based on GS-EDDI data)
- Flood shocks
- Price input shocks
- Price output shocks
- Death of a household member

- Illness of a household member
- Crop production-related shock
- Livestock keeping/production-related shock

Since EDDI (as well as SPEI) data are ideal for capturing droughts but less suitable for measuring floods, for the latter we rely on high-resolution precipitation time series from the Climate Hazards group Infrared Precipitation with Stations (CHIRPS) (Funk et al., 2015), which we use to construct the Standardized Precipitation Index (SPI) and, in turn, build a flood shock dummy—constructed in the same manner as for droughts (i.e., taking value 1 if the household experienced floods (SPI values below -1) during growing season months for a share of months above the 75th percentile of the distribution, and 0 otherwise). Data for all other covariate and idiosyncratic shocks are based on self-reported information from households. CLARE estimation for this additional exercise follows the same Table 1 pipeline used for drought shocks, with the only difference being the alternative shock variable. Nearly 60 percent of households were exposed to at least one of these shocks (cf. Table A.2 in Appendix A).

4.2 Main results

We start by reporting the results of the empirical application of CLARE for drought resilience on our cross-country harmonized longitudinal dataset for four Sub-Saharan African countries. We start with the forecasting task and move to cross-sectional prediction on held-out countries. The outcome, the shock, and the set of confounders and resilience components are the same for both tasks and have been described in the data section. In addition, we compare CLARE’s out-of-sample performance with that of other alternatives, primarily focusing on the resilience score of Cissé and Barrett (2018), which was identified as the best-performing resilience indicator according to the comprehensive study by Upton et al. (2022). Finally, we demonstrate the robustness of predictive performance across alternative shock measures and definitions, as well as different sets of resilience components. In presenting the results, we begin with those from the binarized version of CLARE, as binary resilience scores are the most intuitive and appealing for policy makers for targeting purposes (Upton et al., 2022).²¹

²¹ Across all the empirical analysis, to maintain polarity without altering the correlation with the outcome, the following standardized variables have been included in the final CLARE aggregation (post-estimation) as their complements: *Age of the hh head (lag)*; *Distance to market (lag)*; *Gender of hh head (lag)*; *Rural (lag)*; *% members female >65 (lag)*; *% members male 16-65 yrs (lag)*; *% members male >65 (lag)*.

4.2.1 Forecasting

In this forecasting test, we trained and tuned our models on a pooled training set comprising survey waves 1 to 3 for Malawi, Nigeria, Tanzania, and Uganda. Correspondingly, the testing set includes the fourth survey wave for Malawi, Nigeria, and Tanzania as well as survey waves 4 through 7 for Uganda. We therefore have a total of 11 waves in the training set and 7 waves in the testing set. The goal is to forecast food (in)security status of households appearing in the future waves of the testing set, based on resilience weights and out-of-sample conditional probabilities estimated only on the earlier waves of the training set. The longitudinal dataset is unbalanced, and we do not need the same households to necessarily appear in both the training and testing set waves. In fact, once the weights and input-to-output relationships have been estimated on the training set, predictions can be made on new data, provided that the data consists of the same set of variables. For the binary version of CLARE, we follow Lee et al. (2025) and use the country-specific median value of our resilience score to discretize the continuous indicator: households with a score above that threshold are defined as resilient, while those with a score below it are defined as non-resilient. For completeness, we also report how binary classifications change when using the 25th percentile as an alternative threshold.²² We start with the results of the regression exercise and then move to the binary indicator.

Appendix B reports the full estimation results concerning the average and heterogeneous effects of drought shocks on wellbeing (Figure B.1), the values of the data-driven resilience weights (Figure B.2), and tests for the accuracy of the causal forest estimates (Table B.1 in Appendix B). Regarding the shock effects, we estimate an average and statistically significant (at the 1% level) increase of 5.3 percentage points in the probability of becoming food insecure following a drought shock,²³ with substantial heterogeneity around this average. Concerning the weights (Figure B.1), the three most influential resilience components—jointly accounting for over 50% of the total weights—are the lagged values of the Food Consumption Score (FCS), the asset index, and distance from the market. These top-ranked variables are closely aligned with the conceptual framework of poverty traps, reflecting well-established patterns such as the critical role of assets, the persistence of wellbeing, the importance of nonlinear welfare dynamics, and the centrality of market proximity (Carter & Barrett, 2006; Dercon, 2004;

²² As a reminder, the discretization of CLARE can be done using any threshold depending on the user's preference; here we use these cutoffs purely for illustrative purposes.

²³ The unconditional probability of experiencing food insecurity is 20.3 percent (see Table A.2 in Appendix A). For comparison, this estimated average effect of droughts aligns with those reported for the continuous FCS by Upton et al. (2022) for Niger and Ethiopia.

Fafchamps, 1992). This alignment reinforces the theoretical foundations of the approach and enhances the credibility of the empirical analysis.

Moving to the evaluation of out-of-sample performance, and starting with the regression task, Table 3 reports Pearson’s pairwise and Spearman rank correlation coefficients between the 0-100 version of the CLARE score and the Food Consumption Score, as well as the ratio between the Root Mean Squared Error (RMSE) of the predicted FCS on the testing set waves, based on a regression of the FCS on CLARE in the training set waves, and the sample mean of the FCS. The latter is a test conducted by Upton et al. (2022) to assess the predictive skills of continuous resilience indicators. It is computed by regressing the FCS on CLARE in the training set, then predicting the outcome on the unseen data of the testing set, and finally computing the testing set RMSE normalized by the average value of the FCS in the testing set sample. We also report the average estimated value of CLARE in the testing set.

The metrics suggest a moderately high correlation between CLARE and future FCS, indicating that the indicator can predict the continuous wellbeing outcome with reasonably high accuracy, which is particularly noteworthy given the complexity of forecasting on a large, cross-country household dataset that includes not only future data points but also many households previously unencountered by the algorithms, relying solely on variables from earlier waves collected years prior²⁴ and quantities estimated from an independent training set.

Table 3: Forecasting the Food Consumption Score out-of-sample using CLARE

– All countries, waves 4 to 7

Metrics	Value
Pearson’s correlation coefficient	0.534
Spearman rank correlation coefficient	0.524
Normalized RMSE (RMSE over FCS sample mean)	0.320
Average CLARE score	33.558

Notes: This table reports the out-of-sample performance of the CLARE model in predicting the continuous Food Consumption Score (FCS) across all countries and waves 4 to 7.

²⁴ See Constenla-Villoslada et al. (2025) for a discussion on the challenges of using ML to forecast future food security outcomes when training data is outdated or not updated.

Turning to the classification task, Table 4 reports the confusion matrix illustrating the out-of-sample accuracy of CLARE in forecasting food insecurity for households in the pooled testing set. As a reminder, food insecurity is a binary variable defined as a Food Consumption Score equal to or below 35. Sensitivity, i.e., the ability to identify true positives, is above 82 percent. This means that CLARE can correctly identify in advance more than 8 out of 10 households in those four countries who are going to be food insecure in future waves. Specificity, the ability to identify true negatives (i.e., food-secure households), is lower, around 57 percent, while total accuracy is 61 percent.

Table 4: Forecasting food insecurity status out-of-sample using CLARE

– All countries, waves 4 to 7; median CLARE cutoff

		Food insecure (FCS \leq 35)		
		Food insecure = 0	Food insecure = 1	Total
Binary CLARE (median cutoff)	Non-resilient = 0	5,788	366	6,154
	Non-resilient = 1	4,427	1,728	6,155
	Total	10,215	2,094	12,309
Correctly predicted		56.7%	82.5%	61.1%

Notes: This table shows the out-of-sample classification performance of the binary CLARE model in forecasting food insecurity (defined as $FCS \leq 35$) across all countries and waves 4 to 7. The classification uses the median CLARE value as the cutoff. Rows indicate the predicted values (based on CLARE), while columns represent the observed food insecurity status.

These numbers are, of course, subject to the threshold we choose to use to discretize the continuous CLARE indicator. This, in turn, depends on the preference of the policy makers regarding minimizing false positives versus false negatives. In this respect, it is important to acknowledge the presence of a trade-off. For instance, selecting a threshold below the median would increase specificity and overall accuracy, but at the cost of reducing sensitivity. This would result in fewer false negatives, yet it would also identify fewer non-resilient households.

There is no single ‘right’ or ‘best’ threshold, as the choice depends on the specific priorities set in policy making. One could argue that the cost of incorrectly classifying a food-insecure household as food-secure is higher than the opposite, i.e., false negatives are costlier than false positives. Under this preference, a resilience-building program based on these considerations would prioritize sensitivity over specificity

and accuracy, as it happens to be the case with the median cutoff, although not by design. In doing so, it would provide preventive support to four-fifths of households at risk of food insecurity in the near future. On the other hand, others might argue that spending public funds to assist households that ultimately are not at risk constitutes an inefficient use of resources. For this reason, users may prefer to classify a smaller share of households as non-resilient—specifically, those scoring below the 25th percentile of the CLARE distribution. Table 5 presents the corresponding confusion matrix. By design, this threshold substantially increases specificity and overall accuracy, reaching levels above 82 percent and 78 percent, respectively. While sensitivity decreases, it still identifies 59 percent of non-resilient households.

Table 5: Forecasting food insecurity status out-of-sample using CLARE
– All countries, waves 4 to 7; 25th percentile CLARE cutoff

		Food insecure (FCS ≤ 35)		
		Food insecure = 0	Food insecure = 1	Total
Binary CLARE (25 th percentile cutoff)	Non-resilient = 0	8,365	866	9,231
	Non-resilient = 1	1,850	1,228	3,078
	Total	10,215	2,094	12,309
Correctly predicted		81.9%	58.6%	77.9%

Notes: This table shows the out-of-sample classification performance of the binary CLARE model in forecasting food insecurity (defined as FCS ≤ 35) across all countries and waves 4 to 7. The classification uses the 25th percentile of CLARE as the cutoff. Rows indicate the predicted values (based on CLARE), while columns represent the observed food insecurity status.

4.2.2 Out-of-country prediction

In the out-of-country tests, we trained and tuned our models on a rotating training set comprising three countries and predicted out-of-sample on the left-out country. The goal is to forecast the food (in)security status of households in a different country, based on resilience weights and out-of-sample conditional probabilities estimated from data in other countries. For the binary version of CLARE, we use the country-specific median value of resilience capacity to discretize the continuous indicator. Households with a score above that threshold are defined as resilient, while those with a score below it are defined as non-resilient. Note that since we are doing cross-sectional prediction here—an exercise policy makers might be interested in for transfer learning in data-scarce environments where updated panels are not promptly available—unlike in the forecasting exercise, the rotating testing set of each country includes

all waves of data for that country, and so does the training set for the other countries.

The cross-country results are shown in Tables 6 and 7. For the regression task, Table 6 the performance metrics of the continuous CLARE score are even better than those of the forecasting exercise, with out-of-country correlation coefficients approaching 0.6.²⁵ For the classification task (Table 7), the results are similar to those of the forecasting exercise: specificity around 58 percent, sensitivity 83 percent, and accuracy above 63 percent.

Table 6: Predicting the Food Consumption Score out-of-country using CLARE – All countries

Metrics	Value
Pearson’s correlation coefficient	0.567
Spearman rank correlation coefficient	0.562
Normalized RMSE (RMSE over FCS sample mean)	0.320
Average CLARE score	52.990

Notes: This table reports the out-of-country performance of the CLARE model in predicting the continuous Food Consumption Score (FCS) across all countries.

Table 7: Predicting food insecurity status out-of-country using CLARE

– All countries; median CLARE cutoff

		Food insecure (FCS ≤ 35)		
		Food insecure = 0	Food insecure = 1	Total
Binary CLARE (median cutoff)	Non-resilient = 0	13,089	967	14,056
	Non-resilient = 1	9,315	4,741	14,056
	Total	22,404	5,708	28,112
Correctly predicted		58.4%	83.1%	63.4%

Notes: This table presents the out-of-country classification performance of the binary CLARE model in predicting food insecurity (FCS ≤ 35) across all countries, using the median CLARE value as the cutoff. Rows show predicted values (based on CLARE), while columns indicate observed food insecurity status.

²⁵ The average CLARE score is significantly higher compared to that of the forecasting exercise. This is because the FCS increases over time in the sample, so when the model is trained on all waves from certain countries, rather than only on earlier waves, it results in higher values of the estimated quantities for the held-out countries. In any case, CLARE is a *relative* resilience indicator, with estimated values tailored to the training set, making comparisons across different data inappropriate.

While not impressive in absolute terms, it is important to contextualize them concerning the complexity of the task at hand. The idea is that even if only a single cross-sectional dataset containing the necessary information to construct the 18 resilience subcomponents is available in a given country, and if the policy maker prioritizes maximizing sensitivity at the cost of increased false positives, a CLARE model trained on data from entirely different countries will still accurately identify over 83 percent of households at risk of food insecurity in the future.²⁶

4.3 Comparison with other resilience measures and alternative approaches

What matters in terms of performance is not the absolute metrics *per se*, but how CLARE compares to existing methods. The above results are not directly comparable to the performances reported in other papers on resilience measures, given the different (much larger and more heterogeneous) sample we employ, as well as the outcome used to measure wellbeing and the set of resilience characteristics.²⁷ We, therefore, replicate the predictive exercises of the previous sections using three other measures: the resilience score of Cissé and Barrett (2018), the naïve approach of using the lagged food insecurity or FCS value to predict future status, and the realized resilience measure employed by RIMA and TANGO. For the Cissé and Barrett (2018) indicator, we repeat both the forecasting and the out-of-country exercise. For the other two resilience measures, we only report the comparison for the forecasting exercise, but not for the out-of-country prediction. Finally, we show the value added of CLARE compared to potential alternative and simpler estimation approaches, including food insecurity prediction using standard supervised ML, as well as the construction of the composite indicator based on a simple average with equal weights across resilience subcomponents. To facilitate a clearer comparison, the main results are summarized in Figures 1 and 2 towards the end of this section, including the summary performance of the CLARE metrics shown in the tables above. The full set of results is provided in the Online Appendix.²⁸

4.3.1 The Cissé and Barrett (2018) Resilience Score

We replicate the out-of-sample exercise conducted by Upton et al. (2022) based on the Cissé and Barrett (2018) moments-based method, which—as anticipated in Section 2—provides a robust framework for

²⁶ For space constraints, from now, we will only report classification results using the median cutoff. Results with the alternative 25th cutoff for out-of-country prediction are, however, similar to those of the forecasting exercise.

²⁷ For instance, since consumption-related measures are considerably more volatile, and thus difficult to predict, compared to asset-based measures (Barriga-Cabanillas et al., 2025; Lee et al., 2025), our performance metrics cannot be compared with those from Cissé and Barrett's (2018) application, which used Tropical Livestock Units as the outcome variable.

²⁸ The Online Appendix can be accessed here: <https://shorturl.at/fvP3I>.

measuring resilience as a normative condition. This method, grounded in the conceptual framework developed by Barrett and Conostas (2014), defines resilience as the probability of an individual or household remaining above a minimum normative standard of wellbeing—such as a poverty line, food consumption score (FCS), or nutritional threshold—even when exposed to stressors and shocks. Unlike static measures of wellbeing, the Cissé and Barrett method integrates stochastic elements to capture the dynamics of resilience, making it particularly suited for policy and program evaluations.

The Cissé and Barrett method employs a conditional moment approach to estimate resilience as a probabilistic measure, relying on a two-stage econometric modeling process. In the first stage, the conditional mean of a wellbeing indicator—in our case the FCS—is estimated through an Ordinary Least Squares (OLS) regression. The dependent variable is regressed on the lagged value of the wellbeing indicator, including higher-order terms to account for non-linear dynamics, along with time-varying household and community characteristics and shocks. The squared residuals from this regression are then employed in a second-stage regression to model the conditional variance of the wellbeing indicator, using the same set of covariates as the first equation. These two equations provide the conditional mean and variance of the wellbeing indicator, which are subsequently combined under an assumed two-parameter probability distribution—in this case the gamma distribution—to derive the household-specific probability density function. From this, the resilience score (RS) is calculated as the probability of a household’s wellbeing exceeding the normative threshold of 35 for the FCS. The resulting resilience score, originally ranging from 0 to 1, is converted to a percentage by multiplying it by 100.²⁹ A binary variable is then created to distinguish between resilient and non-resilient households. Similar to CLARE, different thresholds can be employed to discretize the indicator. To ensure consistency between indicators, we use the country-specific median value to discretize the resilience score. Households with a resilience score at or above the median are classified as resilient, while those with a score below the median are considered non-resilient.

We present the results using the same metrics previously employed for CLARE and drawn from Upton et al. (2022). The results of the continuous test, shown in summary Figure 1, reveal a considerably weaker correlation between the FCS and C&B RS (Pearson’s and Spearman’s coefficients) and a higher normalized RMSE,³⁰ indicating lower overall predictive accuracy for the continuous version of this indicator compared to CLARE. The results for the binary indicators suggest that CLARE outperforms

²⁹ For more details on the estimation of the Cissé and Barrett’s (2018) resilience score, see Appendix B.

³⁰ A smaller RMSE indicates a stronger explanatory power of the resilience indicator.

the Cissé and Barrett method across all evaluated metrics (see summary Figure 2 and the full confusion matrix in Table C.1 in the Online Appendix³¹). Overall, while the differences with the C&B RS for the continuous test are substantial, they are not as large for the classification exercise, likely due to the loss of variation and information associated with the discretization of the indicators.

Moving to the out-of-country prediction, we adapt the procedure employed by Upton et al. (2022) for cross-sectional prediction, implementing the Cissé and Barrett (2018) approach. They estimated the resilience score using 75 percent of a country’s household sample and then predicted the remaining 25 percent out-of-sample. Similarly, we estimate the C&B resilience score (RS) using a rotating training sample composed of three countries and then predict the resilience score for the held-out country. We aggregate all out-of-country resilience scores and compare them with the observed wellbeing status for the same households. Table C.2 in the Online Appendix reports the predictive performance metrics for the continuous outcome, while Table C.3 presents the confusion matrix for the binary measure. Similar to the forecasting comparison, both tables show that this model underperforms relative to CLARE, exhibiting lower classification metrics and a higher normalized RMSE. The gap is particularly pronounced in terms of correlation coefficients.

4.3.2 The ‘naïve’ approach

We also evaluate the simplest approach for forecasting future food insecurity, which consists of using the lagged value of the binary food insecurity variable as a forecast for subsequent observations. This method leverages the temporal continuity in food security data, assuming that recent food security status provides valuable information about future conditions. Upton et al. (2022) highlight that even the best-performing resilience measures, such as the C&B RS, shown in the previous section, often fail to consistently improve upon this straightforward approach. Building on these findings, we now compare our measure, CLARE, to this baseline to assess its predictive performance and added value and present the result in the aggregated Figure 1 (for regression) and 2 (classification).³²

Looking at Figure 1, we observe that CLARE substantially outperforms the naïve approach across all examined regression metrics. In line with the findings of Upton et al. (2022), this does not hold for Cissé and Barrett’s RS, which shows predictive power comparable to using lagged wellbeing to forecast future

³¹ The second-to-last wave corresponds to wave 3 for all countries with the corresponding FCS values of wave 4. However, given the availability of additional waves for Uganda, the analysis is extended to include the RS and FCS pairs across waves 4 and 5, 5 and 6, and 6 and 7. Figure 1 presents the mean C&B RS values derived from all iterations of this analysis.

³² The comprehensive results of the classification exercise are reported in Table C.4 in the Online Appendix.

wellbeing. For classification (Figure 2), the results clearly depend on the rule used to discretize CLARE. While the lagged variable approach outperforms binary CLARE in terms of overall accuracy when using the median cutoff for binarization, it severely underperforms in terms of sensitivity, that is, its ability to identify ex-ante food-insecure households, which is the focus of our evaluation. This lower sensitivity means the approach is more likely to misclassify food-insecure households as food-secure. In other words, because most households in the testing set are food-secure, the naïve approach faces a rare-event issue, leading it to overpredict the majority class to boost overall accuracy. However, this comes at the cost of severely misclassifying the minority class, resulting in sensitivity worse than a random guess (34.7 percent). By contrast, the binary CLARE version built using the 25th percentile cutoff achieves overall accuracy comparable to the naïve indicator, while delivering substantially higher sensitivity.

4.3.3 The ‘realized resilience’ approach (RIMA’s and TANGO’s binary indicator)

The third and final comparison with other existing resilience metrics involves the concept of "realized lack of resilience", defined as negative changes in observed food insecurity status over time, as an additional test of household resilience. This approach mirrors the binary resilience indicators employed by FAO’s Resilience Index Measurement and Analysis (RIMA) model and TANGO International. While their resilience approaches are based on the construction of composite continuous indicators, they often require binary classification in practice to identify households as resilient or not. According to the FAO (2016) guidelines, RIMA is not typically used for binary classification of households as food secure/insecure or poor/non-poor, presenting challenges in evaluating progress toward resilience-building objectives. When binary classification is necessary, FAO recommends the use of “realized resilience”, defined as a non-negative change in the wellbeing indicator of interest over successive periods in longitudinal data. Similarly, TANGO’s approach constructs a Resilience Capacity Index (RCI) based on latent variables estimated through factor analysis of diverse indicators. However, when binary classification is required, TANGO also adopts the realized resilience criterion.

Building on this concept, for the classification task, we employ the realized lack of resilience from the previous wave, which is a dummy variable taking the value 1 for lagged negative changes in household FCS across two waves, and zero otherwise, to forecast food insecurity in the current wave. For the regression task, we use the continuous change in the Food Consumption Score (FCS) from the previous wave. As shown in Figures 1 and 2, and in the detailed confusion matrix in Table C.5, this indicator performs poorly out-of-sample across both regression and classification tasks and is outperformed by all other indicators we evaluated.

4.3.4 Prediction of food insecurity with supervised machine learning

What is the added value of our causal ML approach compared to standard predictive models of food insecurity, which are increasingly common in the literature, and have also been used to predict household resilience (Knippenberg et al., 2019; Villacis et al., 2024)? To assess this, we conduct a classification exercise by developing a simple predictive model of food insecurity using random forests (with 1,000 trees) and the set of 18 resilience components included in the construction of CLARE. The results, presented in the aggregated Figure 2,³³ allow for a direct comparison. For consistency with the benchmark CLARE results in Table 4, we apply the median predicted probability as a threshold to categorize households as resilient or non-resilient. The findings indicate that while the random forest model performs comparably to the Cissé and Barrett (2018) binary indicator in terms of out-of-sample accuracy, it consistently underperforms relative to CLARE—when using the median cutoff for CLARE binarization and 50 percent probability for the discretization of the random forest score—across all evaluation metrics.

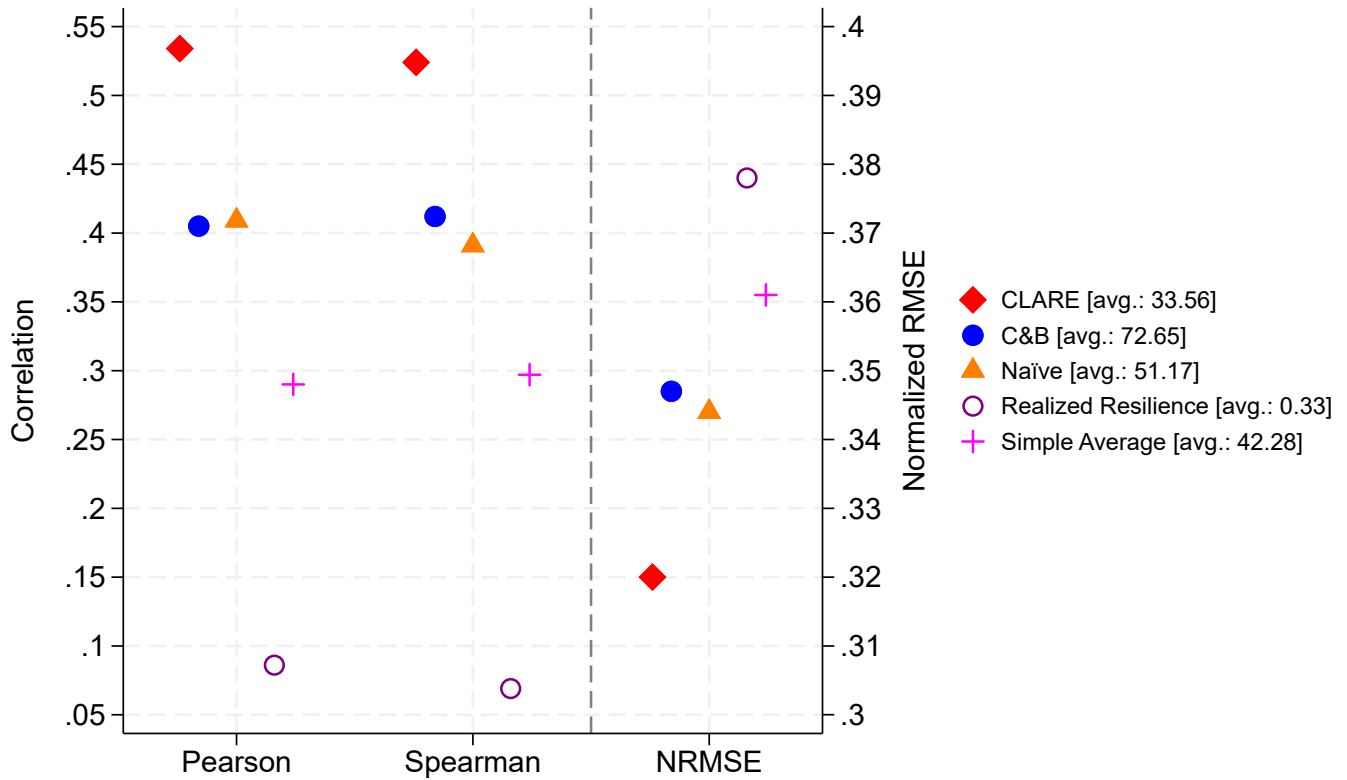
4.3.5 Simple average of resilience subcomponents

Finally, we demonstrate the advantage of using CLARE and its data-driven weighting scheme over a naïve approach to constructing composite indicators, which relies on a simple average of the standardized resilience subcomponents with equal weights (1/18 in our case, reflecting 18 components). In Figure 1, we see that this is the worst-performing measure among all continuous indicators. For the classification task (Figure 2), we see that the simple average indicator—when binarized using the same median cutoff as the benchmark CLARE—demonstrates that merely aggregating the resilience subcomponents is insufficient for achieving strong performance. Instead, aggregation must be guided by criteria that reflect the underlying complex relationships.³⁴

³³ Detailed results are presented in Table C.6 in the Online Appendix.

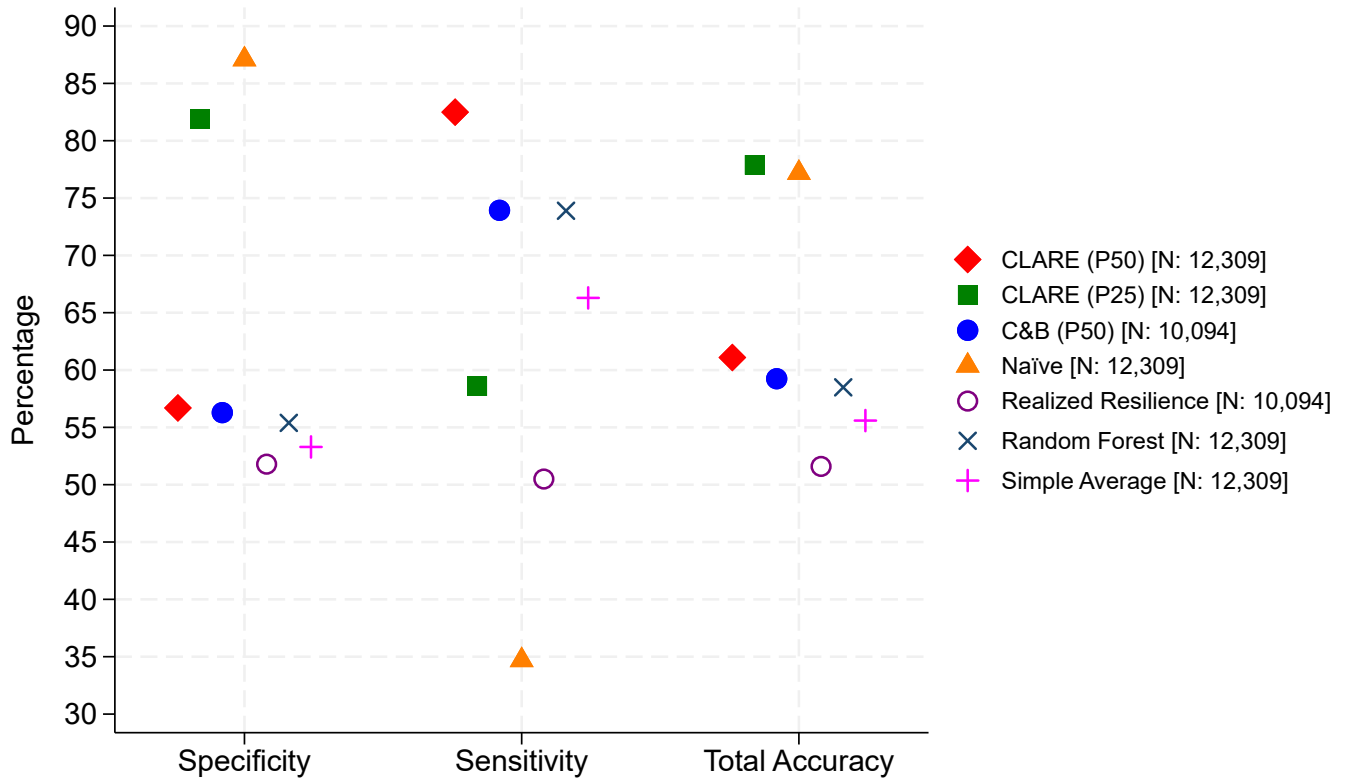
³⁴ Complete results for the classification task are reported in Table C.7 in the Online Appendix.

Figure 1: Summary comparative assessment for the forecasting regression task
 – All countries, waves 4 to 7



Notes: This figure summarizes the out-of-sample performance of the continuous (0-100) CLARE indicator compared to several alternative measures in forecasting the Food Consumption Score (FCS) on unseen future data. The testing set comprises observations from all countries in waves 4 to 7. Reported values correspond to correlations and normalized RMSE (i.e., the same measures used by Upton et al. (2022)); values in parentheses indicate the average measure score.

Figure 2: Summary comparative assessment for the forecasting classification task
 – All countries, waves 4 to 7



Notes: This figure provides a summary comparison of the performance of binary CLARE (using both the median and 25th percentile cutoffs) relative to several alternative measures. The testing set comprises observations from all countries in waves 4 to 7. The number of observations (shown in parentheses) differs across methods due to variations in data requirements and model structures. Detailed results for the alternative measures are reported in the Online Appendix.

4.4 Additional analyses and robustness checks

Having established that CLARE outperforms other approaches, we now assess its robustness and performance stability. For ease of comparison, the main results are summarized in Figure 3 for the forecasting regression task and in Figure 4 for the forecasting classification task. Note that the axis ranges are the same as those in Figures 1 and 2. The insights from the comparative assessment above remain valid irrespective of the specific CLARE version considered relative to alternative approaches. The complete set of results for the binary analysis is reported in the Online Appendix.

4.4.1 Smaller set of resilience components

CLARE enables the computation of resilience in hard-to-reach or data-scarce environments where data

collection is difficult, costly, or both. To ensure its applicability in such contexts, it is necessary that CLARE’s performance does not degrade significantly if estimated using only a relevant subset of key predictors. Figures 3 and 4 show, for the forecasting exercise, the performance metrics when we estimate CLARE using only three raw resilience components pinpointed as having the highest importance weights according to the baseline results, namely, FCS, the asset index, and distance from the market as reported in Figure B.2 in Appendix B. As the reader can see, the results change only slightly for both the continuous and binary³⁵ metrics, indicating that the performance does not degrade significantly. This suggests that CLARE could be effectively estimated in data-poor environments by collecting data on only these key variables without compromising its out-of-sample ability to identify the least resilient households. Furthermore, the model demonstrates stability in performance, even when focusing solely on the top resilience determinants and discarding less significant components. This has practical implications for fieldwork and policy design, suggesting that data collection efforts could be streamlined by prioritizing a core set of variables. This is a particularly valuable characteristic of our indicator, enabling transfer learning in problematic settings. To this end, a single wave of data on a limited set of key variables is sufficient to compute the resilience indicator, thereby offering quantitative support for prioritizing policy interventions toward households more vulnerable to food insecurity in data-scarce environments. Similar reasoning applies to prioritizing data collection efforts in these contexts for out-of-country prediction.

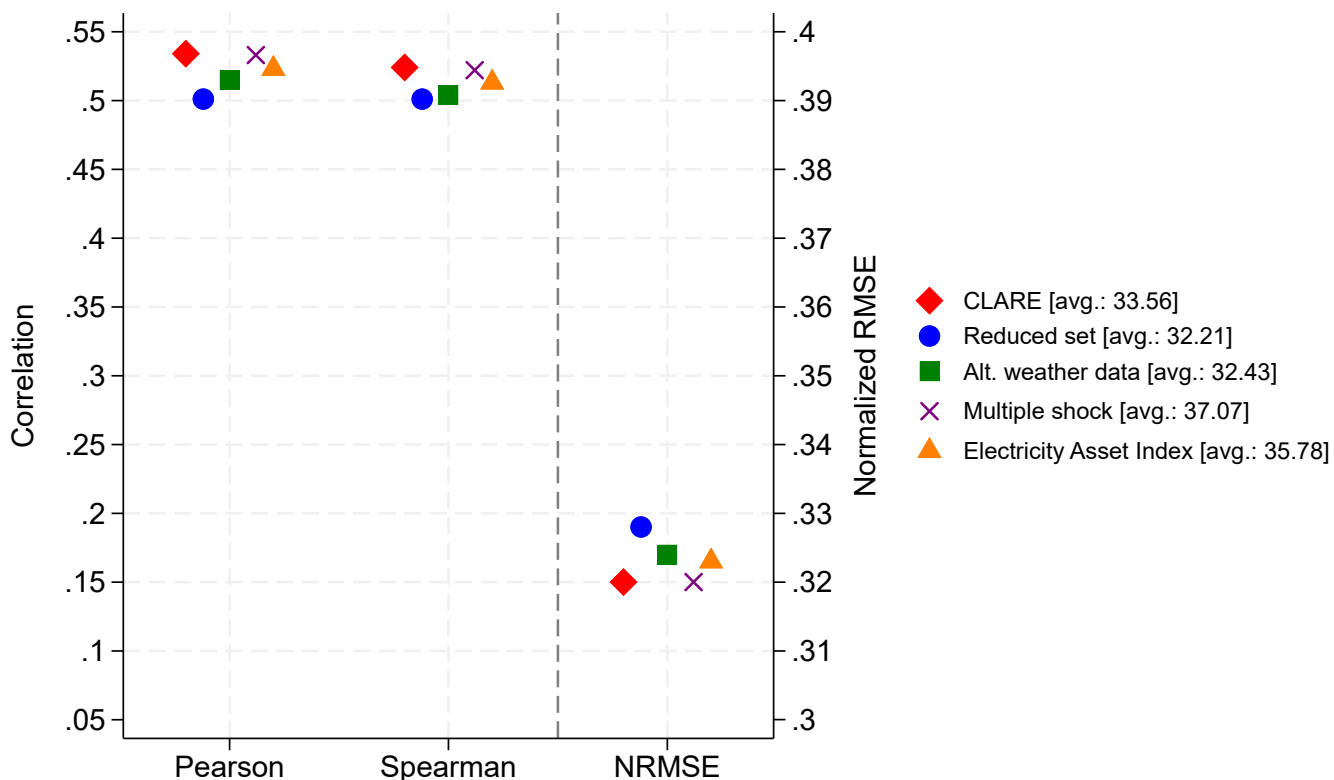
4.4.2 Alternative weather data

We also examine the sensitivity of model performance and the stability of component weights with respect to the source of drought shock data used. Figures 3 and 4 (Table C.9 in the Online Appendix for detailed results) present, among others, the results for the continuous and binary versions of CLARE estimated using drought shocks derived from the SPEI dataset introduced in the data subsection. Figure C.1 in the Online Appendix reports the corresponding variable importance weights. Reassuringly, despite relying on a completely different weather data source to capture drought conditions, both performance metrics and subcomponent weights remain highly consistent, and only slightly inferior to the results reported in Table 3, possibly due to the lower resolution of the weather data employed to compute the shock. Regarding the weights, although there is broader variability, overall consistency remains strong, as the top three and bottom two variables are identical across both importance rankings, albeit in a

³⁵ Table C.8 in the Online Appendix shows the full confusion matrix.

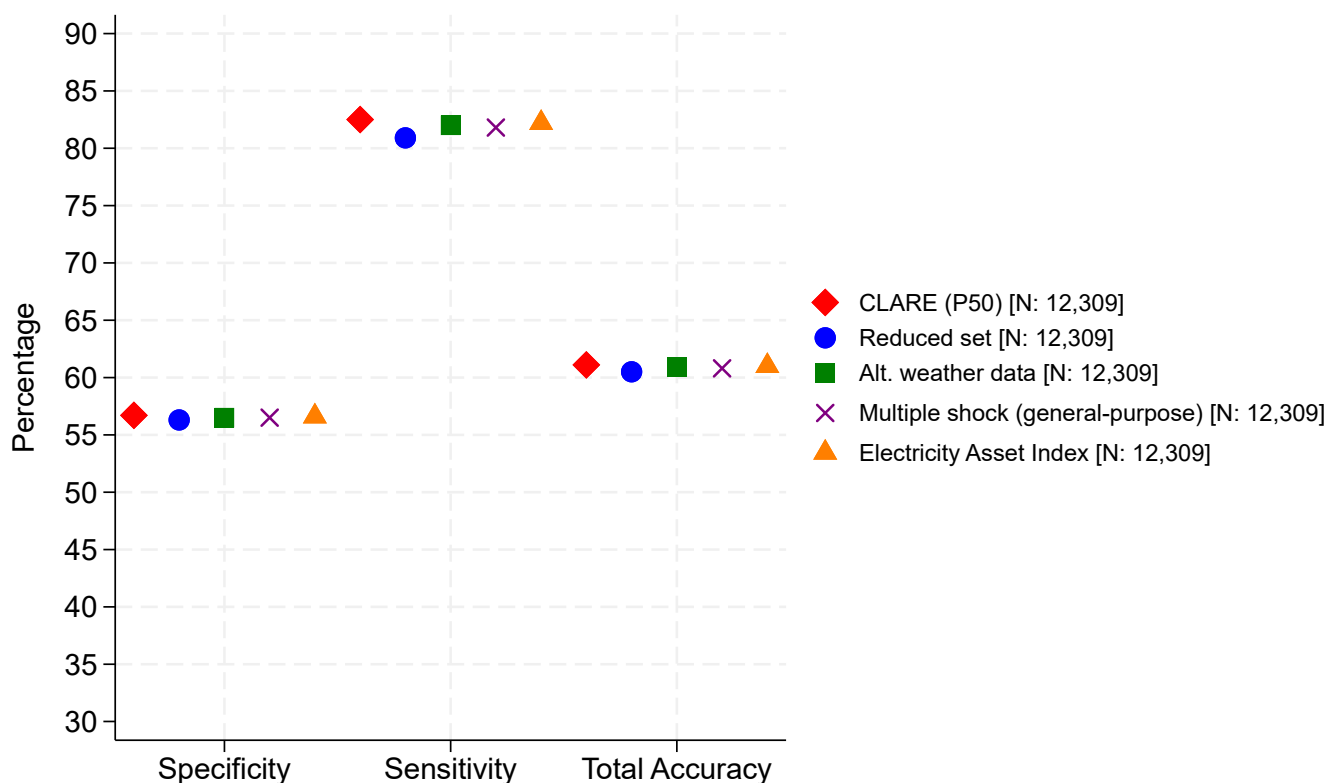
different order.

Figure 3: Summary comparative assessment among different CLARE versions for the forecasting regression task – All countries, waves 4 to 7



Notes: This figure presents performance metrics across different CLARE versions in the forecasting regression task. The testing set comprises observations from all countries in waves 4 to 7. The red diamond labeled ‘CLARE’ in the legend represents the baseline CLARE indicator used for the forecasting analysis. Values in parentheses indicate the average measure score.

Figure 4: Summary comparative assessment among different CLARE versions for the forecasting classification task – All countries, waves 4 to 7



Notes: This figure presents performance metrics across different CLARE versions in the forecasting classification task. The testing set comprises observations from all countries in waves 4 to 7. The red diamond labeled ‘CLARE (P50)’ in the legend represents the baseline CLARE indicator used for the forecasting analysis. For all CLARE versions, the median value was used as the cutoff for binarization of the continuous indicator. The number of observations is in parentheses. The detailed results for each specification are provided in the Online Appendix.

4.4.3 Multiple shock exposure

All the results discussed above pertain to a shock-specific version of CLARE, namely a measure of resilience to droughts. We now present forecasting performance and variable importance for an alternative, general-purpose version of CLARE, constructed based on exposure to a broader set of shocks beyond drought alone (see subsection 4.1 for details on the alternative shock variable). Performance metrics are reported in Figures 3 and 4,³⁶ while the corresponding variable importance ranking is shown in Figure C.2 in the Online Appendix. Again, forecasting performance is fully comparable to that of the drought-specific CLARE, and the variable importance ranking remains very similar. The average shock

³⁶ See Table C.10 for in-depth results on the confusion matrix.

effect is smaller—2.23 percentage points—but statistically significant at the 1 percent level. This attenuated impact likely stems from how the general shock variable is defined: it takes the value of 1 if a household experienced at least one of several possible shocks, thus capturing only a lower bound of the effect associated with multiple-shock exposure. A more thorough investigation of CLARE's estimation under multi-shock frameworks is left to future work.

4.4.4 Alternative asset measure focused on electricity access and energy consumption

Furthermore, we present the results from a robustness check whereby, during the construction of our drought-specific CLARE indicator, we replace the general household asset index with an electricity asset index that includes only items related to electricity access and energy consumption (cf. Table A.4 in the Appendix A for a list of these items).

Resilience frameworks should extend beyond a narrow focus on SDG 1 and SDG 2, shifting the lens from merely measuring end results to recognizing the essential role of enabling conditions. Access to energy—and the ownership of energy-related assets—plays a critical role in building resilience to climate shocks. Energy access enhances the functionality of essential services, supports climate-resilient agriculture, ensures clean water availability, and facilitates education and disaster preparedness. Together, these elements strengthen the capacity of households and communities to adapt to climate stressors (World Bank, 2023).

In the spirit of SDG 7, energy access should be viewed not simply as a sectoral goal, but as a foundational component of stability and continuity for households, services, and local economies during disruptions. A lack of energy access impedes economic recovery, erodes resilience, and undermines progress toward poverty reduction. Conversely, energy access improves quality of life and expands the enabling environment for services, employment, and markets. In particular, the productive use of electricity in rural areas fosters socioeconomic development and income generation, laying the groundwork for community-level resilience and sustainable growth (World Bank et al., 2025). Including such goals refocuses the resilience lens toward the underlying conditions that shape outcomes.

This reasoning motivates the robustness check aimed at assessing whether energy-related assets function as a key transmission channel through which climate shocks affect wellbeing. If confirmed, this would underscore the importance of energy access and consumption-oriented policies and programs as tools to support climate adaptation in rural developing areas. Performance results are reported in Figures 3 and 4 (refer to Table C.11 in the Online Appendix for the full binary results), while Figure C.3 in the Online

Appendix presents the corresponding variable importance ranking. Although regression and classification performance metrics remain qualitatively stable, the household electricity asset index emerges as the second most important driver of the impact of droughts on food security. This suggests that policies targeting the components of this index could strengthen the resilience capacity of households affected by drought—and more broadly, by climate-related stressors.

4.4.5 Summary statistics

Finally, we compare key characteristics of households classified as resilient versus non-resilient, based on whether their resilience score is above or below the median. Table C.12 in the Online Appendix presents the mean values of socio-economic and demographic variables for both groups, along with the statistical significance of the differences between them. We use the baseline CLARE scores as estimated for the full sample in the forecasting exercises. As expected, results reveal notable disparities between resilient and non-resilient households. Resilient households have a significantly higher Food Consumption Score and report lower food insecurity. Resilient households tend to be larger in size and have higher literacy levels, with more members attaining primary, secondary, or higher education. Although both groups are predominantly rural, resilient households display better asset ownership, as indicated by a higher household asset index and similar levels of livestock ownership (TLU). Access to key services within the village is similar across groups, but resilient households are located closer to markets. Importantly, there is no significant difference between the two groups in exposure to the shock—drought conditions measured by the GS-EDDI.

5. Discussion and conclusion

Accurately measuring household resilience is critical for designing, targeting, and evaluating interventions that mitigate vulnerability to shocks, especially in low-income, shock-prone countries. Despite the increasing emphasis on resilience within global development agendas (Béné et al., 2017; Barrett et al., 2021), existing resilience indicators often lack scalability, do not explicitly incorporate the key role of shocks and stressors, and fail to predict wellbeing outcomes accurately in contexts outside their original production environment. This limitation hampers their usefulness in guiding resilience-building policies and interventions (Upton et al., 2022; Barrett & Conostas, 2014). As the demand for resilience-focused interventions continues to grow, particularly in a polycrisis era marked by systemic shocks, pandemics, climate change, and multiple other stressors, there is a pressing need for more robust and scalable methods for resilience estimation (Barrett et al., 2024).

The approach introduced in this paper directly addresses these limitations by employing causal ML techniques to develop a composite resilience indicator capable of overcoming the shortcomings of mainstream methods. CLARE leverages data-driven importance weights for aggregation and conditional probabilities of food (in)security under shocks estimated with flexible causal ML models (Wager & Athey, 2018; Chernozhukov et al., 2018, Chernozhukov et al., 2024), capturing the heterogeneous impacts of shocks across different household subgroups. This allows for a more refined and empirically grounded estimation of resilience, addressing the complex ways in which various subcomponents dampen the effects of adverse events on household wellbeing. Importantly, CLARE is not tied to any specific methodology. As long as the chosen approach estimates granular and heterogeneous treatment effects and allows for establishing an objective hierarchy among the drivers of heterogeneity, any causal ML technique can be employed to identify the underlying causal relationships between shocks, outcomes, and resilience components. Similarly, the data-driven importance of resilience components can be estimated using various model-specific or model-agnostic approaches commonly applied in interpretable machine learning (Lundberg & Lee, 2017; Molnar, 2020).

A significant advantage of CLARE lies in its scalability and flexibility. The use of causal ML enables the estimation and tailoring of variable importance weights based on empirical evidence, facilitating the aggregation of resilience subcomponents in a data-driven manner. This not only improves the accuracy of resilience measurement but also ensures that CLARE can be applied across diverse geographic and temporal contexts. CLARE can be employed to compute resilience capacity indicators that pertain to different specific shocks (e.g., droughts, floods, conflict, market shocks) or a general shock measure incorporating several different shocks and, additionally, can use any possible wellbeing outcome. Finally, it is adaptable to the diverse preferences of policy makers regarding prioritization and funding efficiency. By establishing objective thresholds for prioritizing sensitivity over accuracy, it implicitly performs a cost-benefit analysis of resilience interventions through alternative minimization procedures, which allows for either reducing the number of false negatives to enhance program efficiency or increasing the number of false positives to ensure broader coverage. Overall, thanks to all these characteristics, the adaptability of our approach allows for transfer learning in data-scarce environments and can provide more effective and timely insights for early-warning systems in areas vulnerable to specific shocks, such as climate vulnerability hotspots, thus enabling more targeted and proactive resilience-building interventions.

Our approach, however, is not a panacea. While CLARE addresses many shortcomings of existing

indicators, it does not resolve them all. First, like any other predictive methodology, CLARE is sensitive to non-stationarity issues, known in the ML literature as distribution shifts. A distribution shift occurs when the training data distribution differs from the data distribution the model encounters during testing, causing performance to degrade severely. On this point, recent work by Constenla-Villoslada et al. (2025) shows that using high-frequency monitoring data can greatly reduce non-stationarity issues when forecasting acute child malnutrition with supervised ML techniques. Similarly, the sensitivity of CLARE to this issue might potentially be reduced by integrating standard face-to-face surveys with higher-frequency phone-based surveys or remote sensing data with high temporal resolution.

Second, CLARE relates solely to shocks, not stressors, as Equation 1 is defined only in terms of realizations of a stochastic variable. This issue affects not just CLARE but also other key existing resilience indicators (e.g., Cissé & Barrett, 2018; Alloush & Carter, 2024). CLARE can measure only how sensitive wellbeing is to a shock occurring. But if much of the damage from (uninsured) risk exposure comes through the behavioral adaptations households make to cope with the possibility of shocks, then the counterfactual already accounts for the anticipatory behavioral response to the risk of the shock.³⁷ This, in turn, may lead to an underestimation of the true impact of shocks, part of which affects wellbeing through ex-ante risk mechanisms. One solution to this issue may be to include measures of observed permanent volatility as a proxy for behavioral changes under risk, thereby integrating vulnerability measures into the resilience indicator.

Third, recent work by Lee et al. (2025) demonstrates how sensitive resilience measures can be to the choice of wellbeing indicators used as measurement anchors and proposes several multidimensional alternatives. We did not address this issue here. However, we see their work—focused on the left-hand side of the resilience estimating equation—as complementary to our efforts in this paper to improve resilience estimation on the right-hand side. While CLARE can be applied to any wellbeing outcome, and its weights and performance can, in principle, be shown to be robust and stable across different wellbeing measures (e.g., food security, asset measures, consumption, etc.), this differs from using data-driven methods to fully integrate multiple outcome measures into a single indicator, upon which multidimensional CLARE scores could then be estimated. All these are key avenues for future research.

Finally, the successful operationalization of CLARE fundamentally relies on the availability of multi-

³⁷ We thank Chris Barrett for highlighting this point.

topic, longitudinal and georeferenced survey data. Strengthening existing national longitudinal survey systems in Africa and replicating similar country-owned survey systems elsewhere should be a strategic priority for countries, international organizations and development partners that have stated goals of strengthening resilience. These systems, when paired with innovative resilience measurement approaches like CLARE, have the potential to ensure that resilience-building efforts are both evidence-based and responsive, safeguarding wellbeing and accelerating progress toward sustainable development goals. This need is underscored by recent evidence showing that ML-based forecasting of malnutrition depends on high-frequency survey data (Constenla-Villoslada et al., 2025), and that non-conventional data sources alone, even when combined with ML tools, fail to replicate treatment effect estimates from randomized experiments for key outcomes such as food security (Aiken et al., 2025). However, to more impactfully contribute to the resilience agenda, the next generation survey systems should consider prioritizing (a) leaner and more respondent-centric in-person surveys that are cheaper and disseminate data faster, (b) cross-country harmonization of survey content for resilience measurement, (c) transformation of strictly in-person surveys into mixed-mode operations, increasing adaptability and frequency of data collection through phone interviews between in-person rounds – building on the momentum since the COVID-19 pandemic (Gourlay et al., 2021), and (d) adaptive sampling designs that not only oversample vulnerable households at baseline (a novel approach) but also sample additional households in the future from areas that may be affected by climate, conflict and/or environmental shocks.

References

- Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association*, *105*(490), 493-505.
- Aiken, E. L., Bedoya, G., Blumenstock, J. E., & Coville, A. (2023). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan. *Journal of Development Economics*, *161*, 103016.
- Aiken, E., Bellue, S., Blumenstock, J. E., Karlan, D., & Udry, C. (2025). Estimating impact with surveys versus digital traces: Evidence from randomized cash transfers in Togo. *Journal of Development Economics*, *175*, 103477.
- Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, *603*(7903), 864-870.
- Alkire, S., & Foster, J. (2011). Understandings and misunderstandings of multidimensional poverty measurement. *The Journal of Economic Inequality*, *9*(2), 289-314.
- Alloush, M., & Carter, M. (2024). On the definition and estimation of economic resilience using counterfactuals. *National Bureau of Economic Research (NBER) Working Paper*, No. 33290.
- Arkhangelsky, D., & Imbens, G. W. (2024). Fixed effects and the generalized Mundlak estimator. *Review of Economic Studies*, *91*(5), 2545-2571.
- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353-7360.
- Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685-725.
- Baez, J. E., Kshirsagar, V., & Skoufias, E. (2024). Drought-sensitive targeting and child growth faltering in Southern Africa. *World Development*, *182*, 106702.
- Barrett, C. B., & Carter, M. R. (2013). The economics of poverty traps and persistent poverty: empirical

and policy implications. *The Journal of Development Studies*, 49(7), 976-990.

Barrett, C. B., & Conostas, M. A. (2014). Toward a theory of resilience for international development applications. *Proceedings of the National Academy of Sciences*, 111(40), 14625-14630.

Barrett, C. B., Garg, T., & McBride, L. (2016). Well-being dynamics and poverty traps. *Annual Review of Resource Economics*, 8(1), 303-327.

Barrett, C. B., Ghezzi-Kopel, K., Hoddinott, J., Homami, N., Tennant, E., Upton, J., & Wu, T. (2021). A scoping review of the development resilience literature: Theory, methods and evidence. *World Development*, 146, 105612.

Blumenstock, J. E., Lybbert, T. J., & Putman, D. S. (2025). Probing the limits of mobile phone metadata for poverty prediction and impact evaluation. *Journal of Development Economics*, 103462.

Béné, C., Chowdhury, F. S., Rashid, M., Dhali, S. A., & Jahan, F. (2017). Squaring the circle: Reconciling the need for rigor with the reality on the ground in resilience impact assessment. *World Development*, 97, 212-231.

Béné, C., Newsham, A., Davies, M., Ulrichs, M., & Godfrey-Wood, R. (2014). Review Article: Resilience, poverty and development. *Journal of International Development*, 26(5), 598-623. doi:10.1002/jid.2992.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073-1076.

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

Carr, E. R. (2019). Properties and projects: Reconciling resilience and transformation for adaptation and development. *World Development*, 122, 70–84.

Carter, M. R., & Barrett, C. B. (2006). The economics of poverty traps and persistent poverty: An asset-based approach. *The Journal of Development Studies*, 42(2), 178-199.

Cerqua, A., Letta, M., & Pinto, G. (2025). On the (mis)use of machine learning with panel data. *Oxford Bulletin of Economics and Statistics*, 1-13.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1), C1-C68.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied causal inference powered by ML and AI. *arXiv preprint arXiv:2403.02467*.

Cissé, J. D., & Barrett, C. B. (2018). Estimating development resilience: A conditional moments-based approach. *Journal of Development Economics*, 135, 272-284.

Constas, M., Frankenberger, T., & Hoddinott, J. (2014). Resilience measurement principles: Toward an agenda for measurement design. *Food Security Information Network, Resilience Measurement Technical Working Group, Technical Series, 1*.

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

De Janvry, A., & Sadoulet, E. (2015). Development economics: Theory and practice. *Routledge*.

Dell, M., Jones, B. F., & Olken, B. A. (2014). What do we learn from the weather? The new climate-economy literature. *Journal of Economic Literature*, 52(3), 740-798.

Dercon, S. (2004). Growth and shocks: evidence from rural Ethiopia. In *Macroeconomic policies and poverty* (pp. 308-329). Routledge.

D'Errico, M., & Di Giuseppe, S. (2018). Resilience mobility in Uganda: A dynamic analysis. *World Development*, 104, 78-96.

D'Errico, M., Letta, M., Montalbano, P., & Pietrelli, R. (2019). Resilience thresholds to temperature anomalies: a long-run test for rural Tanzania. *Ecological Economics*, 164, 106365.

Duclos, J. Y., Sahn, D. E., & Younger, S. D. (2006). Robust multidimensional poverty comparisons. *The*

Economic Journal, 116(514), 943-968.

Doan, M. K., Hill, R., Hallegatte, S., Corral Rodas, P. A., Brunckhorst, B. J., Nguyen, M., ... & Naikal, E. G. (2023). Counting People Exposed to, Vulnerable to, or at High Risk from Climate Shocks—A Methodology. *World Bank Policy Research Working Paper*, no. 10619. The World Bank.

Echevin, D., Fotso, G., Bouroubi, Y., Coulombe, H., & Li, Q. (2025). Combining survey and census data for improved poverty prediction using semi-supervised deep learning. *Journal of Development Economics*, 172, 103385.

Fafchamps, M. (1992). Cash crop production, food price volatility, and rural market integration in the Third World. *American Journal of Agricultural Economics*, 74(1), 90-99.

Food and Agriculture Organization (2016). Resilience Index for Measurement and Analysis – II: Analysing resilience for better targeting and action. FAO. Available at:
<https://openknowledge.fao.org/server/api/core/bitstreams/c61a9a8c-feb4-4199-8cb1-7085c84908c8/content>

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., ... & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2(1), 1-21.

Garbero, A., & Letta, M. (2022). Predicting household resilience with machine learning: preliminary cross-country tests. *Empirical Economics*, 63(4), 2057-2070.

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

Harari, M., & Ferrara, E. L. (2018). Conflict, climate, and cells: a disaggregated analysis. *Review of Economics and Statistics*, 100(4), 594-608.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. 2, pp. 1-758). New York: Springer.

Hobbins, M. T., Wood, A., McEvoy, D. J., Huntington, J. L., Morton, C., Anderson, M., & Hain, C. (2016). The evaporative demand drought index. Part I: Linking drought evolution to variations in

evaporative demand. *Journal of Hydrometeorology*, 17(6), 1745-1761.

Hossain, M., Mullally, C., & Asadullah, M. N. (2019). Alternatives to calorie-based indicators of food security: An application of machine learning methods. *Food Policy*, 84, 77-91.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790-794.

Josephson, A., Michler, J. D., Kilic, T., & Murray, S. (2025). The mismeasure of weather: Using Earth Observation data for estimation of socioeconomic outcomes. *Journal of Development Economics*, 103553.

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

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, 25(3), 602-627.

Knippenberg, E., Jensen, N., & Conostas, M. (2019). Quantifying household resilience with high frequency data: Temporal dynamics and methodological options. *World Development*, 121, 1-15.

Knittel, C. R., & Stolper, S. (2025). Using machine learning to target treatment: The case of household energy use. *The Economic Journal*, forthcoming.

Lee, Seungmin; Abay, Kibrom A.; Barrett, Christopher B.; and Hoddinott, John F. (2025). Estimating multidimensional development resilience. *Journal of Development Economics*, forthcoming.

Letta, M., Montalbano, P., & Paolantonio, A. (2024). Climate Immobility Traps: A Household-Level Test. *World Bank Policy Research Working Paper*, no.10724.

Lundberg S., & Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

McBride, L., & Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and

machine learning. *The World Bank Economic Review*, 32(3), 531-550.

Michler, J. D., Josephson, A., Kilic, T., & Murray, S. (2022). Privacy protection, measurement error, and the integration of remote sensing and socioeconomic survey data. *Journal of Development Economics*, 158, 102927.

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu.com.

Monteith, J. L. (1965). Evaporation and environment. In *Symposia of the society for experimental biology* (Vol. 19, pp. 205-234). Cambridge University Press (CUP) Cambridge.

Montalbano, P., & Romano, D. (2022). Vulnerability and resilience to food and nutrition insecurity: A review of the literature towards a unified framework. *Bio-based and Applied Economics*, 11(4), 303-322.

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

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69-85.

Premand, P., & Stoeffler, Q. (2022). Cash transfers, climatic shocks and resilience in the Sahel. *Journal of Environmental Economics and Management*, 116, 102744.

Quandt, A., Neufeldt, H., & McCabe, J. T. (2019). Building livelihood resilience: What role does agroforestry play? *Climate and Development*, 11(6), 485–500.

Ranucci, I., Romano, D., & Tiberti, L. (2025). Weather shocks and resilience to food insecurity: Exploring the role of gender and kinship norms. *World Development*, 188, 106847.

Ratledge, N., Cadamuro, G., de la Cuesta, B., Stigler, M., & Burke, M. (2022). Using machine learning to assess the livelihood impact of electricity access. *Nature*, 611(7936), 491-495.

Scognamillo, A., Song, C., & Ignaciuk, A. (2023). No man is an Island: A spatially explicit approach to measure development resilience. *World Development*, 171, 106358.

Shilpi, F., Kahn, M.E., & Berg, C. 2025. Rethinking Resilience: Adapting to a Changing Climate. Advance Edition. *World Bank Policy Research Report*. Washington, DC: World Bank.

- Smith, L. C., & Frankenberger, T. R. (2018). Does resilience capacity reduce the negative impact of shocks on household food security? Evidence from the 2014 floods in Northern Bangladesh. *World Development*, 102, 358-376.
- Upton, J. B., Cissé, J. D., & Barrett, C. B. (2016). Food security as resilience: reconciling definition and measurement. *Agricultural economics*, 47(S1), 135-147.
- Upton, J., Constenla-Villoslada, S., & Barrett, C. B. (2022). Caveat utilitor: A comparative assessment of resilience measurement approaches. *Journal of Development Economics*, 157, 102873.
- Villacis, A. H., Badruddoza, S., & Mishra, A. K. (2024). A machine learning-based exploration of resilience and food security. *Applied Economic Perspectives and Policy*, Vol.46, 1479-1505.
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-12.
- Walter, I. A., Allen, R. G., Elliott, R., Jensen, M. E., Itenfisu, D., Mecham, B., ... & Martin, D. (2000). ASCE's standardized reference evapotranspiration equation. *Watershed management and operations management*, 2000, 1-11.
- Wiesmann, Doris, Lucy Bassett, Todd Benson, and John Hoddinott (2009). Validation of the World Food Programme's food consumption score and alternative indicators of household food security. *International Food Policy Research Institute discussion paper 00870*.
- Wooldridge, J. M. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*.
- Wolf, S., Twigg, J., Parikh, P., Karaoglou, A., & Cheaib, T. (2016). Towards measurable resilience: A novel framework tool for the assessment of resilience levels in slums. *International Journal of Disaster Risk Reduction*, 19, 280-302.
- World Bank; IEA; IRENA; UNSD; WHO. (2024). *Tracking SDG 7: The Energy Progress Report 2024* (English). Washington, DC: World Bank.
- World Bank (2023). *Transforming lives through energy access in eastern and Southern Africa. World Bank Results Brief*. Washington, DC: World Bank.

Appendix A – Data Details

Table A.1: Variable descriptions

Variable name	Definition
Food Consumption Score (FCS)	Food Consumption Score
Food insecurity	FCS≤35
Share of drought months	Share of growing season months experiencing drought (GS-EDDI > +1)
Drought (GS-EDDI)	=1 if share of drought GS-EDDI > 75th percentile of the country-specific distribution
Multiple shock	= 1 if household experienced any shock ³⁸
Age of hh head	Age of household head
Gender of hh head	=1 if Female household head
Household size	Household size
% members male ≤15 yrs	% members male ≤15 yrs
% members male 16-65 yrs	% members male 16-65 yrs
% members female 16-65 yrs	% members female 16-65 yrs
% members male >65 yrs	% members male >65 yrs
% members female >65 yrs	% members female >65 yrs
Rural	=1 if household lives in a rural area
No. of hh members literate	No. of hh members literate
No. of hh members with primary education	No. of hh members with primary education
No. of hh members with secondary education or higher	No. of hh members with secondary education or higher
Household asset index	Household asset index
Mobile owned	=1 if household own a mobile phone
TLU today	Tropical Livestock Units as of the time of survey
Number of key services within village	No. of key services within community (of 5)
Distance to market	Distance to market (km)

Notes: This table provides definitions for the key variables used in the analysis.

³⁸ The list of shocks included is reported in Section 4.1.4.

Table A.2: Summary statistics – Full sample

Variable	Obs	Mean	Std. Dev.	Min	Max
Food Consumption Score (FCS)	28,112	51.049	19.848	0	126
FCS (lag)	28,112	50.78	19.864	0	126
Food insecurity	28,112	.203	.402	0	1
Food insecurity (lag)	28,112	.21	.408	0	1
Share of drought months	28,112	.086	.093	0	.6
Drought (GS-EDDI)	28,112	.25	.433	0	1
Multiple shock	28,112	.597	.491	0	1
Age of hh head	28,047	49.224	15.278	10	113
Age of hh head (lag)	28,112	47.539	15.537	14	150
Gender of hh head	28,112	.262	.44	0	1
Gender of hh head (lag)	28,112	.245	.43	0	1
Household size	28,112	6.27	3.298	1	35
Household size (lag)	28,112	6.021	3.164	1	35
% members male <=15 yrs	28,112	.19	.173	0	.857
% members male <=15 yrs (lag)	28,112	.2	.179	0	.857
% members male 16-65 yrs	28,112	.23	.196	0	1
% members male 16-65 yrs (lag)	28,112	.241	.204	0	1
% members female 16-65 yrs	28,112	.247	.171	0	1
% members female 16-65 yrs (lag)	28,112	.255	.173	0	1
% members male >65 yrs	28,112	.053	.129	0	1
% members male >65 yrs (lag)	28,112	.044	.12	0	1
% members female >65 yrs	28,112	.067	.158	0	1
% members female >65 yrs (lag)	28,112	.052	.142	0	1
Rural	28,112	.737	.441	0	1
Rural (lag)	28,112	.737	.44	0	1
No. of hh members literate	28,099	3.127	2.358	0	22
No. of hh members literate (lag)	28,112	3.003	2.333	0	18
No. of hh members with primary education	28,099	1.006	1.282	0	13
No. of hh members with primary education (lag)	28,112	1.181	1.445	0	15
No. of hh members with secondary education or higher	28,099	.851	1.387	0	13
No. of hh members with secondary education or higher (lag)	28,112	.798	1.322	0	15
Household asset index	28,112	-.018	.978	-1.23	8.532
Household asset index (lag)	28,112	-.012	.96	-1.174	8.107
Mobile owned	28,094	.69	.462	0	1
Mobile owned (lag)	28,112	.668	.471	0	1
TLU today	27,719	1.203	5.607	0	480.9
TLU today (lag)	28,112	1.253	7.740	0	490.260
Number of key services within village	27,261	3.632	1.509	0	5
Number of key services within village (lag)	28,112	3.668	1.473	0	5
Distance to market	24,899	36.965	37.914	0	214.36
Distance to market (lag)	28,112	36.118	38.149	0	214.36

Notes: This table reports descriptive statistics for the variables used in the analysis, including both current and lagged values. Summary measures are based on the full pooled dataset across countries and waves.

Table A.3: Household asset index components

Malawi	Nigeria	Tanzania	Uganda
Mortar/pestle	furniture (3/4-piece sofa set)	Radio and Radio Cassette	House
Bed	furniture (chairs)	Telephone(landline)	Other buildings
Table	furniture (tables)	Telephone(mobile)	Land
Chair	mattress	Refrigerator or freezer	Furniture/furnishings
Fan	bed	Sewing Machine	Household appliances
Air conditioner	mat	Television	Television
Radio ('wireless')	sewing machine	Video / DVD	Radio/Cassette
Tape or CD/DVD player; HiFi	gas cooker	Chairs	Generators
Television	stove (electric)	Sofas	Solar panel/electric inverters
VCR	stove gas (table)	Tables	Bicycle
Sewing machine	stove (kerosene)	Watches	Motorcycle
Kerosene/paraffin stove	fridge	Beds	Motor vehicle
Electric or gas stove; hot plate	freezer	Cupboards, chest-of- drawers, boxes, wardrobes, bookcases	Boat
Refrigerator	air conditioner	Lanterns	Other Transport
Washing machine	washing machine	Computer	Jewellery and Watches
Bicycle	electric clothes dryer	Cooking pots, Cups, other kitchen utensils	Mobile phone
Motorcycle/scooter	bicycle	Mosquito net	Computer
Car	motorbike	Iron (Charcoal or electric)	Internet Access
Minibus	cars and other vehicles	Electric/gas stove	Other electronic equipment
Lorry	generator	Other stove	Other household assets (e.g., lawn mower)
Beer-brewing drum	fan	Water-heater	
Upholstered chair, sofa set	radio	Record/cassette player, tape recorder	
Coffee table (for sitting room)	cassette recorder	Complete music system	
Cupboard, drawers, bureau	hi-fi (sound system)	Books (not schoolbooks)	
Lantern (paraffin)	microwave	Motor Vehicles	
Desk	iron	Motorcycle	
Clock	tv set	Bicycle	
Iron (for pressing clothes)	computer	Carts	
Computer equipment & accessories	DVD player	Animal Cart	
Satellite dish	satellite dish	Boat/canoe	
Solar panel	musical instrument	Wheelbarrow	
Generator		House(s)	
		Air-conditioned	
		Dish antenna/decoder	

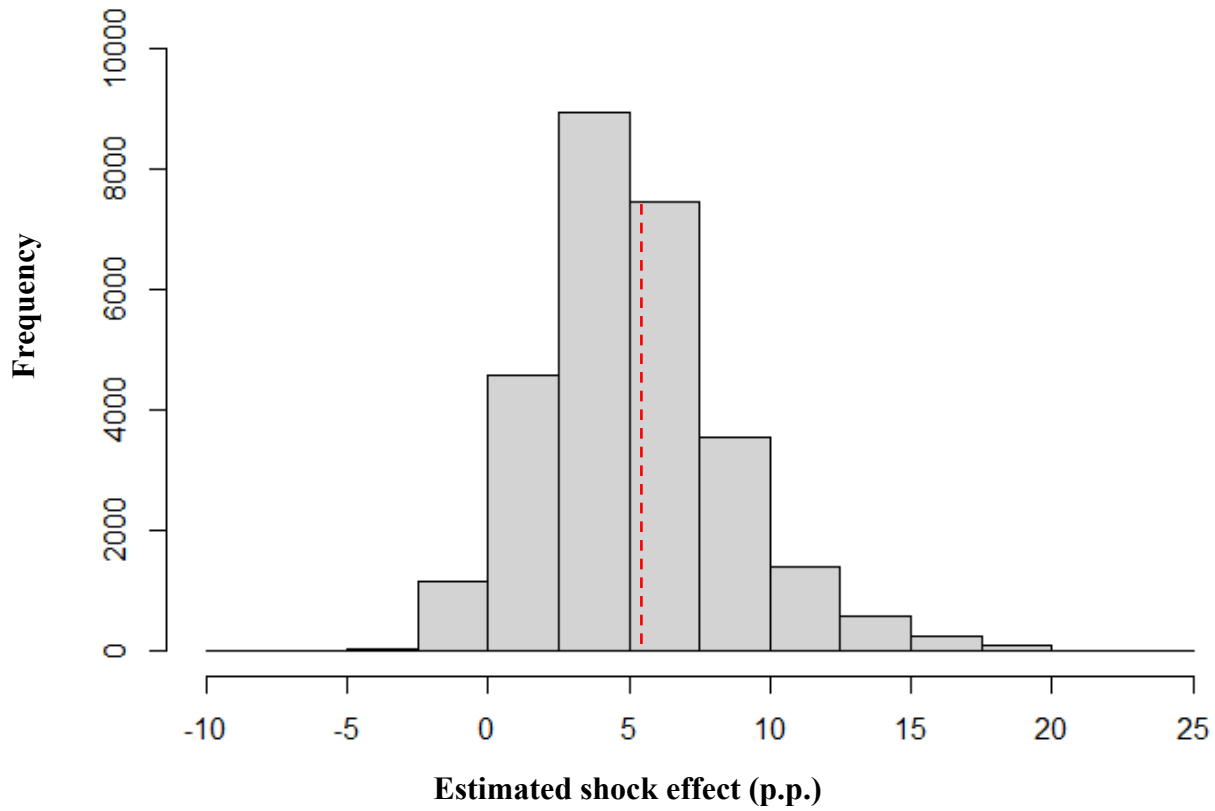
Table A.4: Household electricity asset index components

Malawi	Nigeria	Tanzania	Uganda
Fan	stove (electric)	Radio and Radio Cassette	Television
Air conditioner	stove gas (table)	Telephone(landline)	Radio/Cassette
Tape or CD/DVD player; HiFi	stove (kerosene)	Telephone(mobile)	Generators
Television	fridge	Refrigerator or freezer	Solar panel/electric inverters
VCR	freezer	Sewing Machine	Mobile phone
Sewing machine	air conditioner	Television	Computer
Electric or gas stove; hot plate	washing machine	Video / DVD	Internet Access
Refrigerator	electric clothes dryer	Computer	Other electronic equipment
Washing machine	bicycle	Iron (Charcoal or electric)	Other household assets (e.g., lawn mower)
Iron (for pressing clothes)	motorbike	Electric/gas stove	
Computer equipment & accessories	cars and other vehicles	Water-heater	
Satellite dish	generator	Record/cassette player, tape recorder	
Generator	fan	Complete music system	
	radio	Air-conditioned	
	cassette recorder	Dish antenna/decoder	
	hi-fi (sound system)		
	microwave		
	iron		
	tv set		
	computer		
	DVD player		
	satellite dish		
	musical instrument		

Appendix B – Methodological Annex

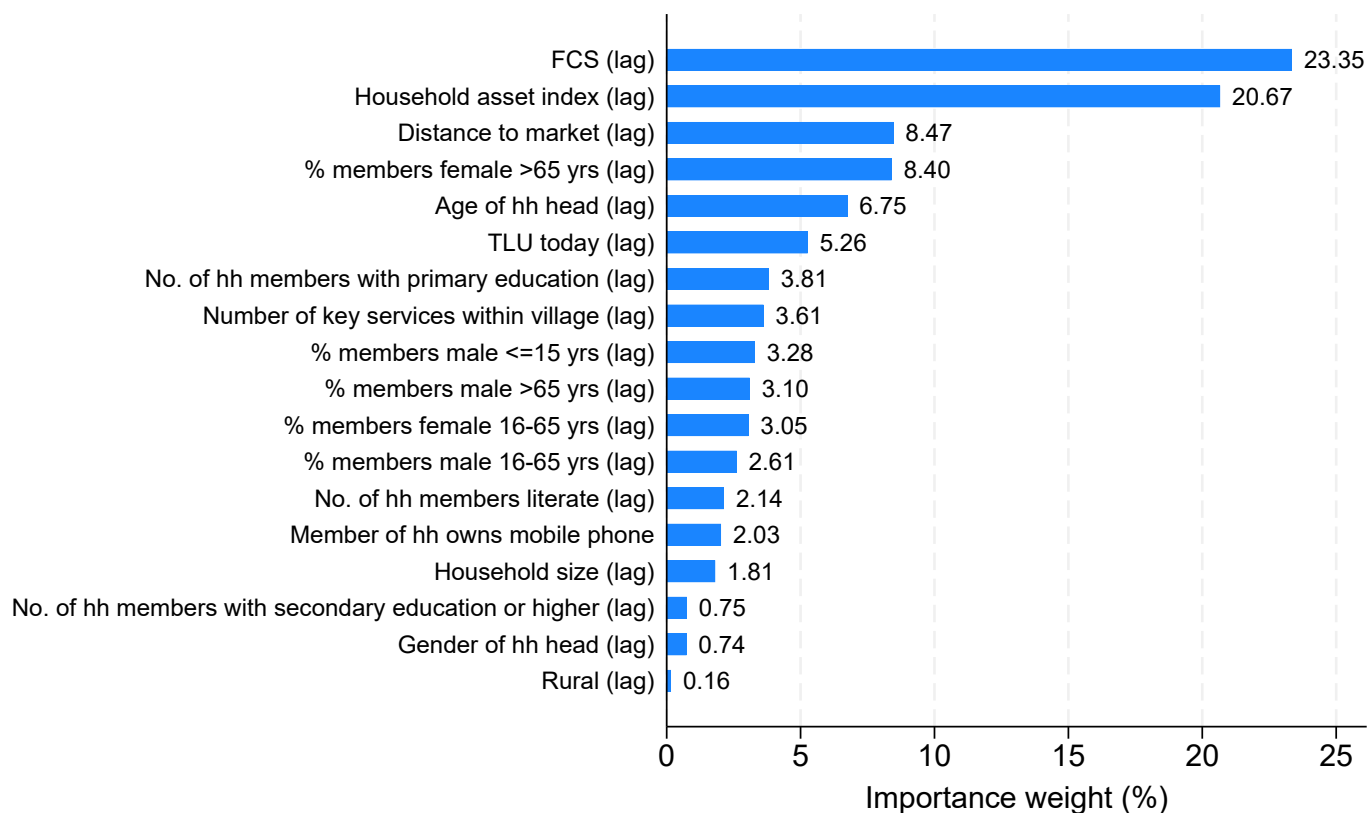
- Causal forest results for the forecasting problem

Figure B.1: Shock effect distribution



Notes: the vertical dotted red line corresponds to the ATE.

Figure B.2: Variable importance weights for the forecasting task



Notes: This figure reports the variable importance weights for the main forecasting task, estimated using causal forests. To maintain polarity without altering the correlation with the outcome, the following standardized variables have been included in the final CLARE aggregation (post-estimation) as their complements: *Age of the hh head (lag)*; *Distance to market (lag)*; *Gender of hh head (lag)*; *Rural (lag)*; *% members female >65 (lag)*; *% members male 16-65 yrs (lag)*; *% members male >65 (lag)*.

Table B.1: Causal forest fit

Calibration test	Estimate	Standard error	t-value	Pr(>t)
Mean forest prediction	0.999	0.149	6.690	0.000***
Differential forest prediction	0.995	0.286	3.473	0.000***

Notes: These are the results of the calibration tests recommended by Athey et al. (2019) to assess the quality of the causal forest fit. A coefficient of 1 for 'Mean forest prediction' suggests that the mean forest prediction is correct, whereas a coefficient of 1 for 'Differential forest prediction' additionally suggests that the heterogeneity estimates from the forest are well calibrated. See [here](#) for more details. Finally, the excess error is below 7.8×10^{-6} .

Methodological Details on Estimating Cissé and Barrett’s (2018) Resilience Score

The Cissé and Barrett (2018) (C&B) method operationalizes the theory of resilience as a normative condition developed by Cissé and Costas (2014), and it refers to the concept of maintaining a sufficiently high probability of exceeding a critical wellbeing threshold. The C&B method leverages econometric techniques to estimate the conditional moments of wellbeing indicators. Below, we provide a detailed breakdown of the steps involved in the estimation of our resilience score.

In this analysis, we employ the same set of variables used in the CLARE estimation. However, in this case, all variables are measured at time t rather than lagged at $t-1$, to align with the original methodology. To ensure the sample size remains consistent across all observations, we handle missing values at time t by imputing them using the past values of the respective variables. The overall imputation rate is approximately 1.32 percent. Given this low percentage, the impact of imputation on the dataset is minimal. While we acknowledge that this imputation approach may not be the most methodologically rigorous, it enables us to maintain a comprehensive dataset for the estimation of resilience scores following the Cissé and Barrett method.

To conduct the out-of-sample exercise, we estimate the Cissé and Barrett resilience score using the first three waves of the dataset. We then assess the predictive performance by comparing the binary classification of resilience scores (RS) derived from wave 3 with the food insecurity outcomes ($FCS \leq 35$) observed in wave 4. For Uganda—due to the availability of a longer panel—this exercise is extended and iteratively estimated across additional wave pairs, specifically waves 4 and 5, 5 and 6, and 6 and 7.

The first step involves estimating the conditional mean of the FCS using an Ordinary Least Squares (OLS) regression:

$$FCS_{it} = \sum_k \beta_k FCS_{i,t-1}^k + \gamma_1 X_{it} + \Omega S_{it} + \varepsilon_{it} \quad (\text{B.1})$$

Where the lagged wellbeing indicator ($FCS_{i,t-1}^k$) captures the dynamic nature of wellbeing while the superscript k denotes the polynomial order. X_{it} includes households’ characteristics and community-level variables, i.e., our resilience components; S_{it} is the same binary shock indicator we use for CLARE estimation, and the residual term (ε_{it}) represents the error term.

The squared residuals from the first-stage regression are then used as the dependent variable in a second-stage regression to estimate the conditional variance as follows:

$$\varepsilon_{it}^2 = \sum_k \alpha_k FCS_{it-1}^k + \delta_1 X_{it} + \vartheta S_{it} + \omega_{it} \quad (\text{B.2})$$

In the third step, using the conditional mean and variance estimates, and assuming a two-parameter probability distribution, in this case the gamma distribution, the probability density function of wellbeing is derived. The conditional probability of exceeding the normative threshold (35) is calculated as:

$$RS_{it} = P(FCS_{it} > 35 \mid FCS_{it-1}, X_{it}, S_{it}) = F(35, FCS_{it}, X_{it}, S_{it}) \quad (\text{B.3})$$

Where $F(\cdot)$ is the complementary cumulative distribution function that determines the probability of exceeding the threshold. The resulting resilience score (RS) is expressed as a probability ranging from 0 to 1 and is scaled to a percentage format for interpretation and comparison with the CLARE method.

To conduct the predictive accuracy test using confusion matrices, a binary classification is required to distinguish between resilient and non-resilient households. To ensure consistency with CLARE, resilient households are defined as those with a resilience score (RS) at or above the median of the RS distribution, while non-resilient households are those with a score below the median. The predictive accuracy test requires a balanced panel, meaning that only households observed in both the last and the second-to-last wave are included. This requirement ensures consistency in the dataset over time, but it also results in a smaller sample size compared to the CLARE method.

Online Appendix

The Online Appendix for this paper is available at: <https://shorturl.at/fvP3l>.