PNAS RESEARCH ARTICLE | ECONOMIC SCIENCES SUSTAINABILITY SCIENCE OPEN ACCESS

High-frequency monitoring enables machine learning–based forecasting of acute child malnutrition for early warning

Susana Constenla-Villoslada[a,b], Yanyan Liu[b,c], Linden McBride[d], Clinton Ouma[e], Nelson Mutanda[e], and Christopher B. Barrett[c,f,1]

Affiliations are included on p. 11.

Contributed by Christopher B. Barrett; received August 9, 2024; accepted April 28, 2025; reviewed by Kathy Baylis, Thomas Heckelei, and Yiqun Xie

The number of acutely food insecure people worldwide has doubled since 2017, increasing demand for early warning systems (EWS) that can predict food emergencies. Advances in computational methods, and the growing availability of near-real time remote sensing data, suggest that big data approaches might help meet this need. But such models have thus far exhibited low predictive skill with respect to subpopulation-level acute malnutrition indicators. We explore whether updating training data with high frequency monitoring of the predictand can help improve machine learning models' predictive performance with respect to child acute malnutrition by directly learning the dynamic determinants of rapidly evolving acute malnutrition crises. We combine supervised machine learning methods and remotely sensed feature sets with time series child anthropometric data from EWS' sentinel sites to generate accurate forecasts of acute malnutrition at operationally meaningful time horizons. These advances can enhance intertemporal and geographic targeting of humanitarian response to impending food emergencies that otherwise have unacceptably high case fatality rates.

food security | food crises | humanitarian response | nonstationarity | resilience

Significance

As the number and share of people suffering from food insecurity worldwide has risen over the past decade, the humanitarian response community increasingly seeks advances in early warning systems to target populations who need food assistance. Advances in Earth Observation data and in machine learning have excited interest in their potential to help with early warning and geographic targeting of food assistance. To date, however, the predictive performance of such models with respect to child acute malnutrition has disappointed. We show how predictive skill and predictors vary over time and demonstrate the high value of monthly monitoring of child anthropometry in sentinel sites. With such data it is feasible to generate reasonably accurate forecasts at time horizons of 6 mo.

Increasingly frequent and severe shocks—whether extreme weather events such as droughts, or socioeconomic ones like conflict or price spikes—have precipitated growing food emergencies worldwide. In 2022, an estimated 349 million people experienced acute food insecurity, more than double the 2017 figure, and the largest number recorded in over 70 y (1). In the absence of timely, well-targeted, responses to such emergencies, people become severely malnourished and more vulnerable to infectious disease, leading to avoidable death. Young children (under 60 mo of age, U5) are especially vulnerable to temporary disruptions in access to healthy foods. Malnourished children lose weight rapidly, resulting in severe acute malnutrition (SAM), or severe wasting, the most serious manifestation of acute malnutrition. SAM is associated with 45% of all worldwide deaths of children under 5 y of age (2, 3). Anthropometric indicators, such as weight-for-height, are the gold standard for establishing SAM, especially among U5 children and at subnational levels (4).

Humanitarian and governmental agencies have invested heavily in early warning systems (EWS) to inform needs assessments, humanitarian appeals, and emergency response precisely to avert avoidable death, especially among young children. Ideally, EWS provide the international community with spatially precise alerts of impending food emergencies that might manifest in a spike in the prevalence of global acute malnutrition (GAM), a category that supplements SAM with moderately acutely malnourished—or "wasted"—children. World Health Organization (WHO) official guidelines allow for the identification of critical intervention thresholds based on a community's GAM prevalence (5). EWS commonly combine high-frequency monitoring of sentinel sites with secondary weather, price, and socioeconomic data that may provide leading indicators of food emergency episodes. The most widespread EWS food emergency prediction initiative, the Integrated Food Security Phase Classification (IPC) (6–10), has been externally validated as informative of impending food crises at a large spatial scale (11–13).[*] But the IPC's coarse spatial scale and lack of information on specific vulnerable populations makes it less well-suited for guiding relief operations to the communities with the highest GAM prevalence in an affected region (9, 10, 14, 15). Employing advances in computing methods and relatively new high frequency data sources, researchers have begun using "big data" approaches to malnutrition indicators

[*] The IPC measures in (6–8) are sourced from FEWS NET, while (9, 10) use data from the IPC system. The former is a widely used input into the latter.

Author contributions: S.C.-V., Y.L., L.M., C.O., N.M., and C.B.B. designed research; S.C.-V. performed research; S.C.-V. analyzed data; C.O. and N.M. contributed data and background information; C.B.B. supervised overall project; and S.C.-V., Y.L., L.M., and C.B.B. wrote the paper.

Reviewers: K.B., University of California, Santa Barbara; T.H., University of Bonn; and Y.X., University of Maryland.

Competing interest statement: Mutanda and Ouma work for the Government of Kenya's National Drought Management Authority, which collected the data and advises the government and others in the humanitarian community on food emergency response.

PNAS policy is to publish maps as provided by the authors.

[1] To whom correspondence may be addressed. Email: chris.barrett@cornell.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2416161122/-/DCSupplemental.

Published June 6, 2025.

at higher spatial resolution and subpopulation levels. Such models, commonly trained on repeated cross-sectional or low-frequency longitudinal (panel) data available through the Demographic and Health Surveys (DHS) or Living Standards Measurement Study (LSMS) (16, 17), have so far exhibited good out-of-sample (OOS) predictive skill for wealth-based poverty indicators (18–23). To date, however, these methods and models have not produced accurate predictions of anthropometric indicators of GAM; not in contemporaneous predictive exercises, and much less in forecasts of future conditions (7, 15, 19, 20, 24). Existing models routinely use anthropometric indicators from DHS or LSMS datasets as the outcome variables to train and test machine learning models. But DHS are repeated cross-sectional surveys conducted only every several years at best, and even panel LSMS surveys only revisit households every several years. Resampling of survey enumeration areas and households, combined with randomized offsets to the georeferenced communities—so as to preserve survey respondent anonymity—makes it difficult to tap any time series information to inform GAM prediction. Consequently, analysts rely mainly on cross-sectional and spatial variation in predictors to identify variation in anthropometric outcomes in infrequently collected data.

We hypothesize that machine learning models' disappointing performance to date in predicting malnutrition outcomes stems not from any shortcoming in the computational methods nor in the feature sets used. Rather, it arises from the intrinsic nonstationarity of malnutrition indicators, especially those associated with acute (rather than chronic) malnutrition. Shocks of the sort that precipitate food emergencies quickly, significantly, and perhaps quite locally, shift the conditional and label distribution of acute malnutrition indicators. Due to distribution shift, machine learning models trained on data with stochastic dynamic generating processes struggle with OOS generalization. So it is perhaps little surprise that models trained on dated or cross-sectional data cannot accurately predict post-shock-conditions. Studies that have used higher frequency longitudinal data of the sort collected by EWS have produced better estimates (15, 25–29). Intuitively, the inclusion of recent observations of the predictand in training data enables the algorithm to update to the new data generating process, enhancing predictive skill. Training machine learning models on longitudinal data collected at a frequency that matches the outcome's real-world distribution shift patterns might therefore enable researchers to tap the full potential of big data approaches for predicting GAM.

In this paper, we show that time series child anthropometric data allows machine learning forecasting of GAM prevalence at the community level. We harness a rich time series available from a long-standing EWS operated by the National Drought Management Authority (NDMA) of Kenya in the nation's arid and semiarid counties, where food emergencies occur regularly. NDMA field enumerators collect monthly household survey data in 23 counties, with data statistically representative at the subcounty administrative level 3 (ADM3; see *SI Appendix*, Fig. S1 for a map of the wards included in the analysis). Within the monthly survey, enumerators use a fixed protocol and equipment to measure mid-upper-arm circumference (MUAC), a widely used anthropometric indicator of GAM, among a rotating panel of randomly selected U5 children. Randomly selected households are repeatedly surveyed for multiple months, but inconsistent household tracking makes household-level analysis unreliable. The local household samples are sufficiently large, however, to yield good estimates of ward-level conditions, as shown in an earlier generation of these data (25). We use 3,616,129 child-and-month-specific MUAC observations from January 2006 to December 2020 to generate ward-and-month specific U5 GAM prevalence, yielding a total of 21,643 observations. GAM is defined per WHO standards as MUAC<125 mm. Such prevalence estimates at this ADM3 spatial scale are commonly used in needs assessments and in geographic targeting of emergency response by government agencies and humanitarian relief organizations in the context of impending malnutrition crises (5).

Our results show that supervised machine learning methods applied to high frequency (monthly) sentinel site anthropometric data, supplemented by readily available, near-real-time remote sensing, conflict, and food price data, generate GAM forecasts multiple months into the future with far greater predictive skill than has been found thus far in the literature. Further, we exploit the unusually long time series available in the NDMA data to demonstrate four key points. First, predictive skill depends heavily on autocorrelation as a measure of data quality and distribution stationarity in the outcome variable. Disruptions in data collection—such as by changing sentinel sites—therefore degrade predictive skill. Second, we show that although the time series is crucial, conditioning on remote sensing feature sets improves the sensitivity and precision of GAM predictions over those obtained by exclusive reliance on longitudinal observations of GAM prevalence. Third, we show how much the variable importance of distinct feature groups varies over time. Variables that are most predictive of episodes of high GAM are not those with the highest unconditional importance across all periods, most of which are not crises. This implies analysts should exercise caution in interpreting or relying on prediction methods that pool across periods without regard to the time series sequencing of observations. Last, oversampling rare episodes of high GAM prevalence in model training boosts predictive skill, but only in the presence of adequate autocorrelation in GAM, and at the expense of a loss in precision, likely due to the difficulty of resampling from a noisy distribution.

Overall, this paper provides a rigorous, detailed example of the underlying data conditions that allow for accurate machine learning forecasting of community-scale wasting prevalence up to six months in advance, and reinforces the case for ongoing longitudinal monitoring of sentinel sites to equip EWS to harness the potential of big data approaches to generate such forecasts (16, 17, 30–32).

Results

We first test the use of the simplest models that a policy maker or other stakeholder might use in this scenario. We estimate a one-lag model, including just the previous month's observation to predict the following month ward value. Such a model relies only on the simplest time series autocorrelation of ward-level wasting prevalence. We also estimate a model with only ward fixed effects, which exploits only the time-invariant spatial variation in wasting prevalence among wards. For comparability purposes, we use just the same amount of data we use in the main application, that is a sliding window of 36 mo of data. As shown in *SI Appendix*, Fig. S2*A*, relying only on spatial variation is totally uninformative for forecasting purposes. *SI Appendix*, Fig. S2*B*, of the same figure shows there is some useful forecasting signal in the time series autocorrelation of the outcome variable, but the model results are quite erratic over time. The results from our main ML-based specifications are far superior. This shows how, first, predictive performance increases and stabilizes with the use of secondary variables that add additional predictive power in this context, and, second, how the use of forest ensemble methods with a greater number of lags increases predictive skill relative to either a simple autorregressive or a purely spatial approach.

We estimate three different machine learning models that differ solely in predictors. The first uses remote sensing and other secondary data (SD model). The second relies exclusively on the time series properties of the ward-specific prevalence measures and incorporates just three lags of the dependent variable [we label this the time-series only (TSO) model]. The third aims to make use of the information gain from both the time series and secondary data (hybrid model). We generate monthly forecasts for five different time horizons: 1, 3, 6, 9, and 12 mo in the future over the 14 y time span covered by the data.

Our paper emphasizes the role of using adequate data for the predictive task at hand. Frequency of data collection is not the only important factor in this aspect, and the long temporal coverage of our dataset allows us an uncommon opportunity to investigate how changes in underlying data patterns in the outcome variable affect model performance. To do so, we divide the whole sample period into four subperiods based on underlying time series patterns of GAM prevalence. This pattern variability reflects changes in either the evolution of wasting prevalence over time, specifically the total number of monthly high prevalence cases in the sample, or widespread disruptions in data collection, namely the change in sentinel site survey locations of 2016. The first subperiod, characterized by a relatively high number of wards with high wasting prevalence, ranges from May 2006 to December 2009. The second subperiod, characterized by a steady decline in the number of wards with high wasting prevalence and in the sample's average wasting prevalence, ranges from January 2010 to June 2016. The third subperiod starts from the sentinel site location change in July 2016 and runs through July 2018, and is characterized by unusually low levels of ward-specific temporal autocorrelation in the outcome variable. The last subperiod begins when autocorrelation returns to the values observed previously in the time series and ranges from September 2018 to December 2020. Vertical dashed lines separate these subperiods visually in the Figures.

We evaluate model performance OOS, where the OOS sample is the nth-month temporal step-ahead, for the full time series as a whole and for each of the particular subperiods, in two ways: by model fit in terms of R^2, and by classification accuracy in terms of sensitivity, precision, and specificity. In the context of an EWS, sensitivity is the most important metric, as it captures the capacity of each model to identify impending, potentially life threatening, malnutrition crises among U5 children. Specificity measures each model's ability to correctly identify negative cases. Precision refers to the percentage of positive predictions that are real crisis events, or true positives, and is therefore related to efficient resource allocation if such forecasts trigger humanitarian aid distribution.

Average Forecasting Performance. Unlike prior studies that struggle to predict child GAM even contempo- raneously in models trained on infrequent anthropometric data, yielding average OOS R^2 between 0.08 and 0.11 (19, 20), our models, trained on monthly data, perform well, even at forecast horizons of 6 mo. The upper panel of Fig. 1 shows the OOS R^2 of each of the three models and for each forecast time horizon. The closest comparison to previous work in the literature, the model trained only on remote sensing and other secondary data [Fig. 1, the SD model in panel (*A*)] yields average OOS R^2 over the 14 y span of the time series of 0.39, 0.29, 0.19, 0.13, and 0.12 for predictions 1, 3, 6, 9, and 12 mo in advance.

The results from the two models that exploit the information contained in the labels' time series—the TSO and hybrid models—significantly outperform the results from the SD model, especially over shorter prediction horizons for which temporal autocorrelation is high. The model relying only on previous values of the time series (Fig. 1*B*)—with no secondary data in the feature set—yields average OOS R^2 of 0.53, 0.32, 0.23, 0.18, and 0.15 for the same prediction horizons. The results from the hybrid model (Fig. 1*C*) yields similar OOS R^2, with average values of 0.53, 0.33, 0.23, 0.19, and 0.15. All models' predictive performance decreases with longer prediction horizons, as expected in a scenario with frequent distribution shifts that cause outdated data to rapidly become uninformative of future states.

This increase in predictive skill from training on high frequency data likely reflects the existence of short- or medium-term temporal autocorrelation in wasting prevalence, shown in Fig. 2*A*. For almost the full time span covered by the data, autocorrelation is high up to 6 mo in advance. That temporal stationarity or short-or medium-term trend continuity can be harnessed for forecasting purposes. The time series information embedded in high frequency—i.e. monthly, which is high frequency for the context—longitudinal data sharply boosts the performance of big data approaches to predicting the future prevalence of malnutrition.

Classification Accuracy. Although R^2 is an informative performance metric, the models are better analyzed in their suitability for early warning and targeting purposes in terms of classification accuracy, that is, how effectively these models can identify communities based on realized GAM prevalence. We follow WHO guidelines and define alert cases that require intervention where wasting prevalence at the community level exceeds 15%. Based on this threshold, we calculate the models' sensitivity, or capacity to detect GAM episodes, along with the model's specificity, or accuracy in identifying non-GAM episodes. We also include precision, the number of positive cases that are true positives, as a measure of resource allocation efficiency. Panel 2 of Fig. 1 shows the average sensitivity and specificity of each one of the models over time. Panel 3 of the same figure shows the average precision results. To ease interpretation, we will focus on the operationally meaningful periods for which performance is best (1-, 3-, and 6-mo in advance). To highlight the incidence of alert cases over time, the figures also show a scatter plot of wasting prevalence values for each month, with gray markers indicating values below the alert threshold of 0.15, and orange markers indicating values above that threshold. Similar figures for the other prediction time horizons (9- and 12-mo in advance) can be found in *SI Appendix*, Fig. S3.

The models' differences in terms of classification accuracy follow patterns similar to the R^2 results. The worst performance corresponds again to the secondary data model, which yields average sensitivity of 0.39, 0.29 and 0.19 for the 1-, 3-, and 6-mo prediction horizons, respectively. Average precision among all monthly predictions in the time series averages 0.41, 0.32, and 0.19 for this same model and time horizons. Last, since positive cases are rare events and negative cases are much more widespread, specificity is generally quite high, with average values of 0.97 to 0.99 in all time horizons considered.

Harnessing the temporal autocorrelation of the outcome variable enhances performance in terms of classification accuracy as well. For the same time-horizons, the time series only and hybrid models yield average sensitivity values of 0.48, 0.33, and 0.20; and 0.53, 0.40, and 0.33, respectively. Precision averages 0.54, 0.38, and 0.23 for the TSO model, and 0.51, 0.32, and 0.18 for the hybrid model. Finally, for both models and in all time horizons considered, average specificity is consistently very high, with values between 0.97 to 0.99. Overall, the hybrid model shows higher sensitivity than the TSO model, this advantage being more pronounced in longer time horizons, at the price of mildly lower precision. This result highlights the usefulness of harnessing remote sensing and other secondary data in forecasting malnutrition.

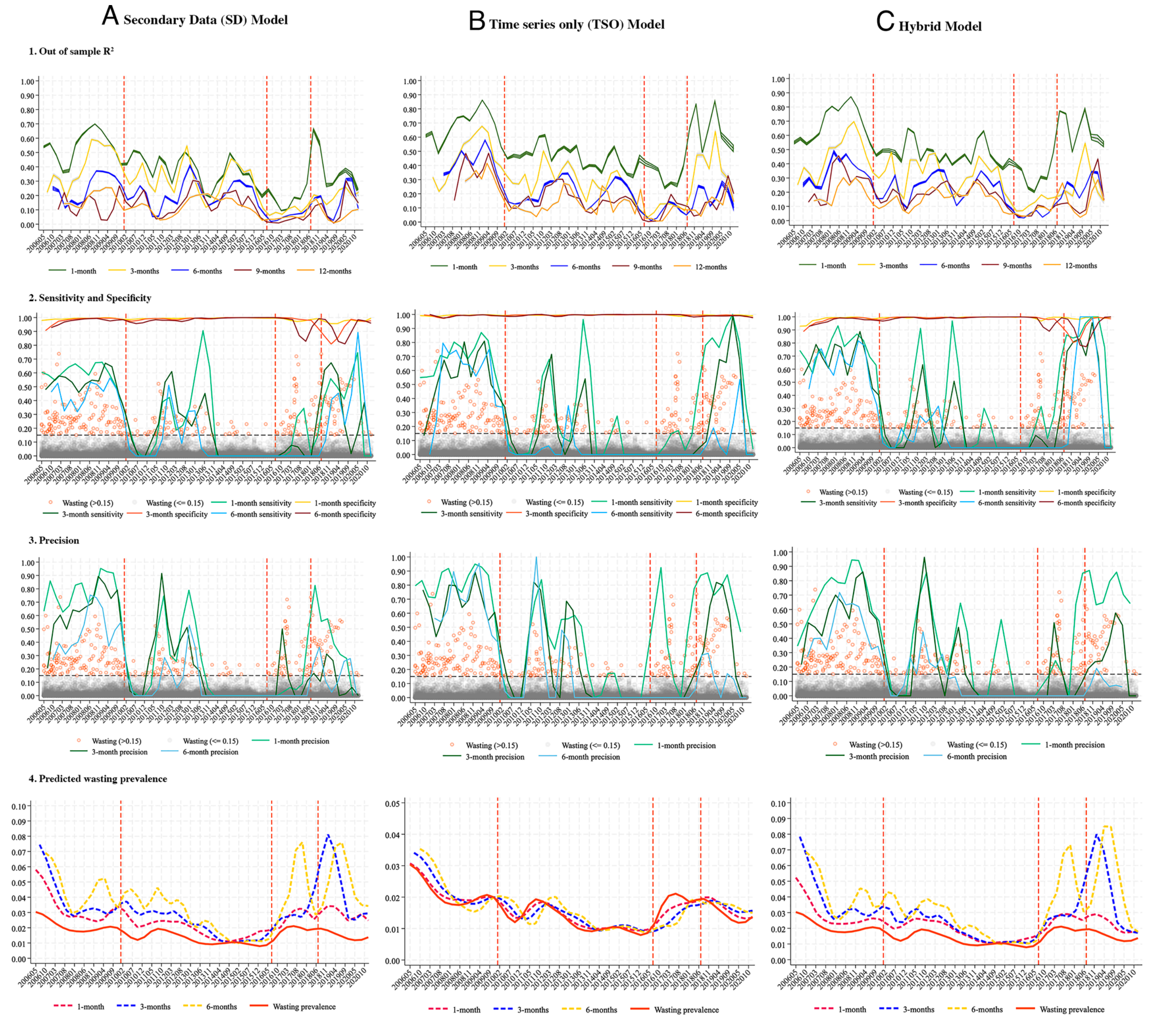
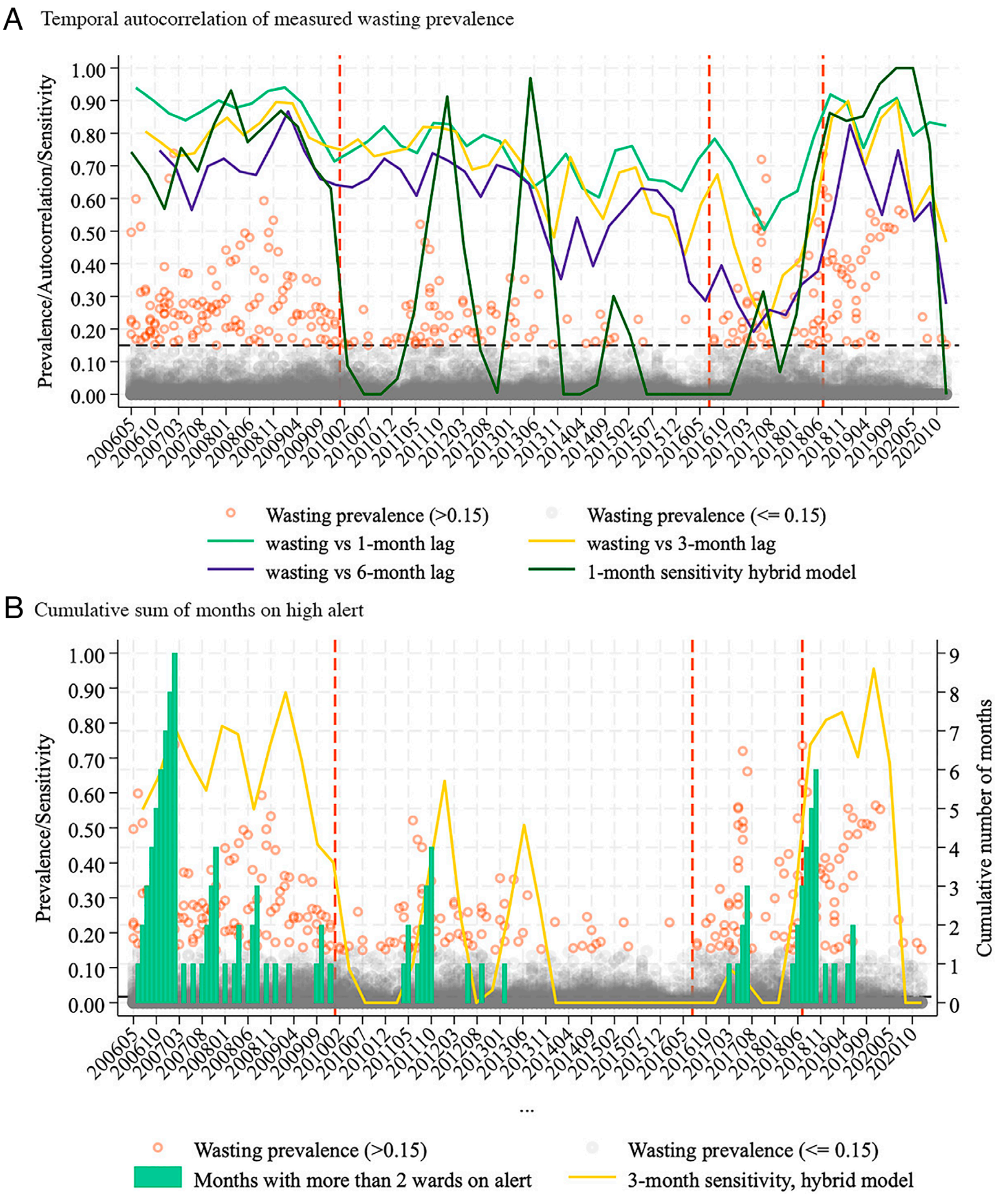


Fig. 1. OOS performance metrics and predicted wasting distributions. The figure shows a local polynomial regression of monthly OOS r-squared (Row 1), sensitivity and specificity (Row 2), precision (Row 3), and predicted wasting prevalence (Row 4), for the SD (*A*), TSO (*B*), and hybrid model (*C*). Predictions are made at 1, 3, 6, 9, and 12 mo in advance. The predicted outcome is ward-level wasting prevalence. The vertical lines in each figure indicate the distinct time periods described in the text. To assist visualization, except in the R^2 case, just the results from wasting predictions that show the best predictive performance—at 1, 3, and 6 mo in advance—are shown. To see how classification accuracy changes according to the underlying distribution, a scatter plot of the measured wasting prevalence is presented in the background of the second and third panels.

Dynamics in Forecasting Performance. The average results obscure considerable dynamic variation in performance due to underlying data patterns. Model performance is strongest during the first subperiod, characterized by the highest average GAM prevalence observed in the time series. Even in this period, ward-level average GAM prevalence values are low overall, with an average value of 0.02 and SD of 0.05, with just 4% of the monthly ward observations (188 out of 4,079) reaching values higher than 0.10 and an average of 3 wards per month getting over the 0.15 threshold. Average OOS R^2 (*SI Appendix*, Fig. SI4, *Upper* panel) for 1 mo in advance predictions in this period equals 0.60, 0.70, and 0.71 for the SD, TSO, and hybrid models, respectively. These average R^2 values have only been achieved previously in nowcasting of much more static malnutrition measures, using household-level indicators such as the Food Consumption Score or Coping Strategies Index (33, 34), but never in predicting anthropometric-based GAM indicators, much less in forecasting them into the future. For the 3-mo prediction horizon, the average OOS R^2 drops to 0.42, 0.48, and 0.48. For predictions 6 mo in advance, the average OOS R^2 drops to 0.26, 0.43, and 0.36, still much better than even nowcasts based on repeated DHS cross-sections (20, 21).

The second subperiod is characterized by a continuous decay in the models' predictive capacities. The subperiod is also characterized by a steady drop in average GAM prevalence, with an average of just 0.01 and a SD of 0.03; just 2% of the observations in this subperiod (175 out of 7,918) registered prevalence greater than 0.10 and an average of one ward per month is over the 0.15 threshold. From January 2010 to June 2016, the average OOS R^2 on the 1 mo in advance predictions equals 0.38, 0.46, 0.48 for the SD, TSO, and hybrid models, respectively. For the 3 mo predictions, the average R^2 for the same models drops to 0.32, 0.29,


A Temporal autocorrelation of measured wasting prevalence

B Cumulative sum of months on high alert

Fig. 2. Autocorrelation and high alert periods. Panel (*A*) shows a local polynomial regression of the temporal autocorrelation of the monthly measured wasting prevalence with its 1-, 3-, and 6-mo temporal lags. Panel (*B*) shows a bar graph of the amount of consecutive months with at least two wards under hunger alert. A local polynomial regression of the 3-mo sensitivity of the hybrid model is also shown. Both panels show a scatter plot of the measured wasting prevalence in the background, with high values above the triggering threshold in orange, and low values in gray.

and 0.32, respectively. Average OOS R^2 for predictions 6 mo in advance equals 0.21, 0.22, and 0.23. This pattern—good prediction when there are many nonzero and larger values of the predictand, weaker when there are fewer nonzero values—is intuitive. In all periods there exists some minimum level of truly random nonzero measures as well as measurement error that is intrinsically difficult to predict. As a larger share of the nonzero observations represent noise rather than signal, predictive skill naturally falls.

The worst model performance occurs during the third subperiod, immediately after the data collection sites changed and coinciding with a big drop in temporal autocorrelation of the outcome variable (Fig. 2*A*). It takes 2 y before model performance recovers. Average OOS R^2 for 1 mo predictions January 2016-July 2018 is 0.16 for the SD model, and 0.31 for the TSO and hybrid models. For the 3 mo prediction horizon, the values go down to 0.09, 0.10, and 0.12. The R^2 of 6 mo predictions is 0.06, 0.07, and 0.05. During this time period, average wasting prevalence equals 0.02 in the sample, with a SD of 0.05, and only 4% of the observations (113 out of 2,279) exceed 0.10 with an average of two wards per month in alert.

After July 2018, model performance starts to recover, coinciding with an increase in the wasting prevalence's temporal autocorrelation to previous values seen in the time series. This performance recovery is particularly evident for the models relying on the time series information as predictors. Once the temporal autocorrelation recovers and the GAM prevalence distribution has higher variability again, performance reaches the high values seen in the first time subperiod. In this time period, the 1 mo average R^2 goes up to 0.41 for the SD model, and to 0.69 and 0.67 for the TSO and hybrid models, respectively. For longer prediction horizons, performance does not recover as quickly. The 3 mo average R^2 equals 0.20, 0.39, and 0.27 for the SD, TSO, and hybrid models, respectively. For predictions 6 mo in advance, average R^2 goes down to 0.15, 0.23, 0.17, respectively. It is important to note that during 2020, the number of observations per month drops drastically, due to data collection issues during the COVID-19 pandemic. Average wasting prevalence during this period is 0.01, with a SD of 0.05, and just 3.5% of the observations (79 out of 2,167) exceeding 0.10 at an average of two wards per month in alert.

To further explore the performance of the three models, we present a local polynomial regression of measured and predicted wasting prevalence values over time in the bottom panel of Fig. 1. Although the hybrid and TSO models have similar R^2 values, the TSO model yields predictions closer to the measured ones. However, the final distributions exhibit a high degree of temporal correlation between the predicted versus the real lagged values on which the model was trained. This could be interpreted as the evolution of the original wasting distribution being purely stochastic, and therefore not predictable further from the periods for which the temporal autocorrelation still holds (35). Consequently, the distributions obtained from predictions from the TSO model look just like temporally displaced copies of the original one. This would also explain the rapid decay in predictive performance of this model over longer prediction horizons. The distributions from the hybrid and SD models, however, do not follow those same patterns. Although more inaccurate in terms of the magnitude of predicted prevalence, the temporal patterns follow the original ones more closely. This difference signals that the secondary data variables add predictive power, which also explains the less drastic performance decay of these models over longer prediction horizons. Our preferred specification is therefore the hybrid model, yielding the best results by combining the information held in the temporal stationarity or autocorrelation of the time series with the forecasting power of the secondary data.

The apparent systematic overprediction of the SD and hybrid models arises due to the truncated nature of the wasting prevalence distribution. Since there are no values below zero and most predictions are very close to zero, the prediction error accumulates on only the positive side of the true value, inflating the average. *SI Appendix*, Fig. S7 shows this; systematic overprediction disappears after removing exact zero-valued observations in the monthly measured prevalence.

Dynamics in Classification Accuracy. As was the case for the OOS R^2, dynamic differences in predictive accuracy over time for all models and performance decay for longer prediction time horizons are evident in the measures of sensitivity, specificity, and precision. Fluctuation patterns in sensitivity are similar to the ones referred to when analyzing model fit results. That is, for all models, there are long periods of very high average performance, followed by others with performance decay. Following the same temporal splitting used to describe the R^2 results, average sensitivity (*SI Appendix*, Fig. SI4, second panel) between May 2006 and December of 2009 for the 1 mo in advance predictions is 0.59, 0.69, and 0.77, for the SD, TSO, and hybrid models, respectively. For predictions 3 mo in advance, the values equal 0.54, 0.54, and 0.68, respectively. For predictions six months in advance, average sensitivity values are 0.44, 0.57, and 0.64. Average precision (*SI Appendix*, Fig. SI4, third panel) is also high during this subperiod, with values of 0.79, 0.82, and 0.72 for the 1 mo predictions of the SD, TSO, and hybrid models, respectively. For predictions 3 mo in advance, the values drop to 0.67, 0.69, and 0.59, respectively. For the 6 mo time horizon, average values equal 0.52, 0.70, and 0.49. Average specificity (*SI Appendix*, Fig. SI4, *Bottom* panel) during this time period is very stable for all models and prediction horizons, ranging between 0.97 to 0.99.

In the subsequent period, from January 2010 to June 2016, alert cases drop, as does overall performance. Sensitivity values for 1 mo predictions during this period drop to 0.23, 0.27, and 0.30, for the SD, TSO, and hybrid models. Precision equals 0.29, 0.37, and 0.36. For 3 mo predictions, average sensitivity values are 0.21, 0.18, and 0.16; and average precision values 0.24, 0.26, and 0.23. In the 6 mo prediction horizon, average sensitivity values drop as low as 0.10, 0.03, and 0.08; and average precision values to 0.09, 0.08, and 0.10. Specificity is still stable at 0.99 for all models and prediction horizons.

During and right after the change in data collection sites, model accuracy is the lowest. Average sensitivity from July 2016 to July 2018 for the 1 mo predictions is 0.10, 0.16, and 0.22 for the SD, TSO, and hybrid models, respectively. Average precision values equal 0.06, 0.35, 0.38, respectively. For the 3 mo prediction horizon, average sensitivity in this subperiod drops to 0.01, 0.03, and 0.03; with average precision values of 0.13, 0.10, and 0.13. Predictions 6 mo in advance during this subperiod record average sensitivities of 0.00 for all models, with also 0.00 average precision. Specificity for all models during this period ranges from 0.96 to 0.99.

In the last subperiod, from July 2018 onward, performance peaks again. One month average sensitivity equals 0.51, 0.82, and 0.89 for the SD, TSO, and hybrid models, with average precision values of 0.52, 0.80, and 0.75. As in the R^2 case, once temporal autocorrelation of the dependent variable is restored, the models recover their predictive skill. For the 3 mo time-horizon, average sensitivity values are still high at 0.43, 0.53, and 0.77, with average precision of 0.14, 0.56, and 0.32. Last, predictions 6 mo in advance register average sensitivity values of 0.35, 0.11, and 0.68 for the SD, TSO, and hybrid models, with average precision values of 0.17, 0.16, and 0.09, respectively. Contrary to previous results, specificity is stable for the TSO model in this period, dropping to its lowest values for predictions with the secondary data and hybrid models. For the TSO model, specificity equals 0.99 for all time-horizons, while the SD and hybrid models average 0.97, 0.91, and 0.88, and 0.99, 0.90, and 0.89 for predictions at 1, 3, and 6 mo in advance, respectively.

The next section explores the underlying data patterns that coincide with these performance fluctuations.

Understanding Dynamic Performance: Serial Autocorrelation and High Alert Periods. Two main contributing factors help explain the drastic changes in model predictive capacity over time. First, we focus on analyzing the common, biggest drop in model performance, during and right after the data collection sites changed in 2016. The survey location change coincides with the onset of a widespread, drought-driven, humanitarian crisis that our data capture but that none of the models forecast accurately.

Since the data collection areas changed during this period and the models are trained on historical data, for a brief period of time, the models are trained with data from different locations. One possible interpretation of the performance drop during this period is that GAM prevalence from one location cannot be predicted accurately with data from other locations. This a priori plausible explanation cannot explain, however, why the drop in performance continues for two more years after the data collection sites changed, especially for the 1 mo in advance predictions of the TSO model, which relies uniquely on the temporal autocorrelation of the outcome measure to harness its predictive power. If the TSO model cannot yield accurate predictions as it did in previous periods, that suggests that there is no temporal continuity in the outcome measure during this period.

Fig. 2*A* sheds light on this hypothesis. It shows a local polynomial regression of the serial autocorrelation of the measured wasting prevalence over time. The correlations shown are of the wasting prevalence at time *t* with its 1- (*t-1*), 3- (*t-3*), and 6- (*t-6*) mo

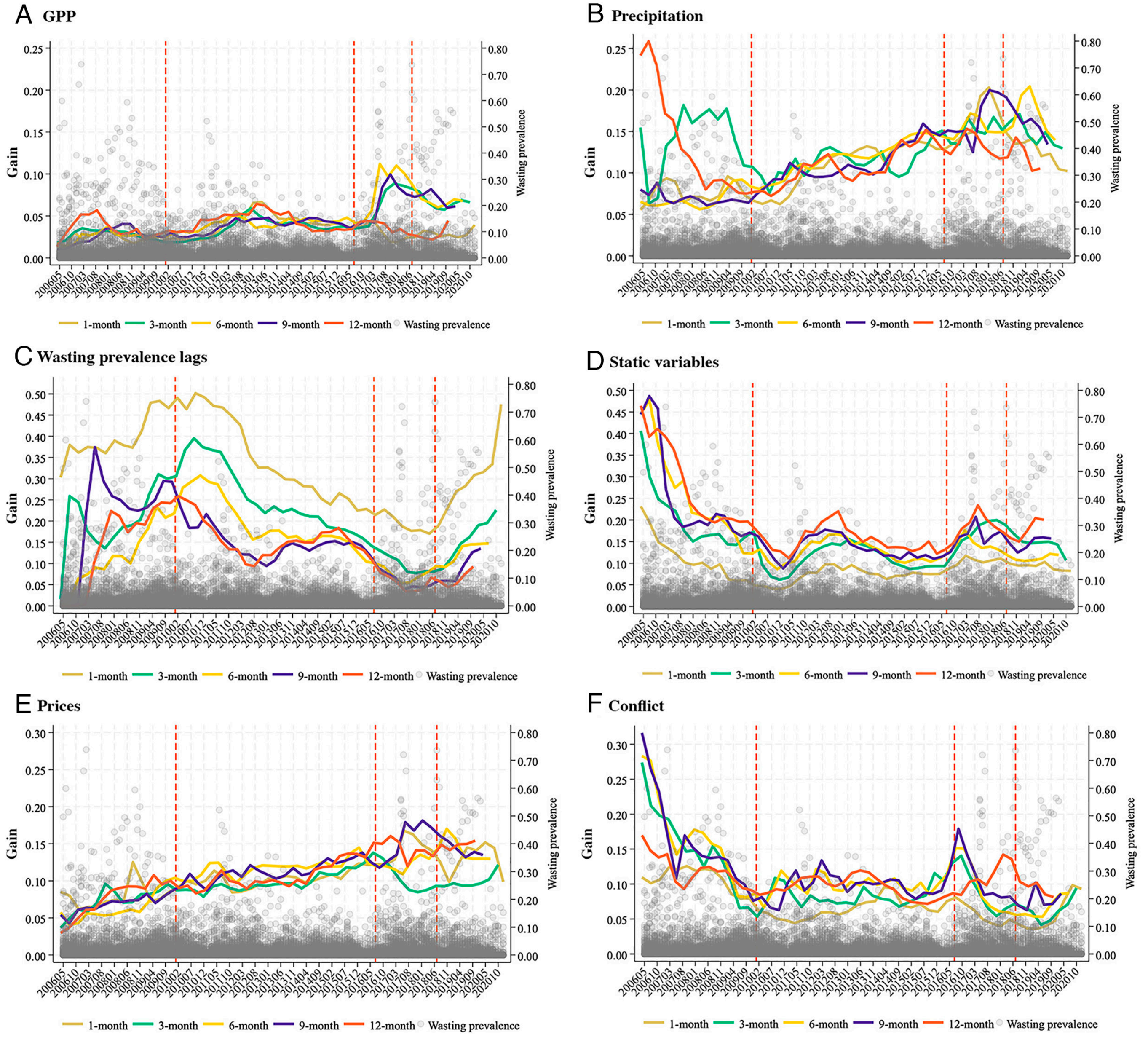


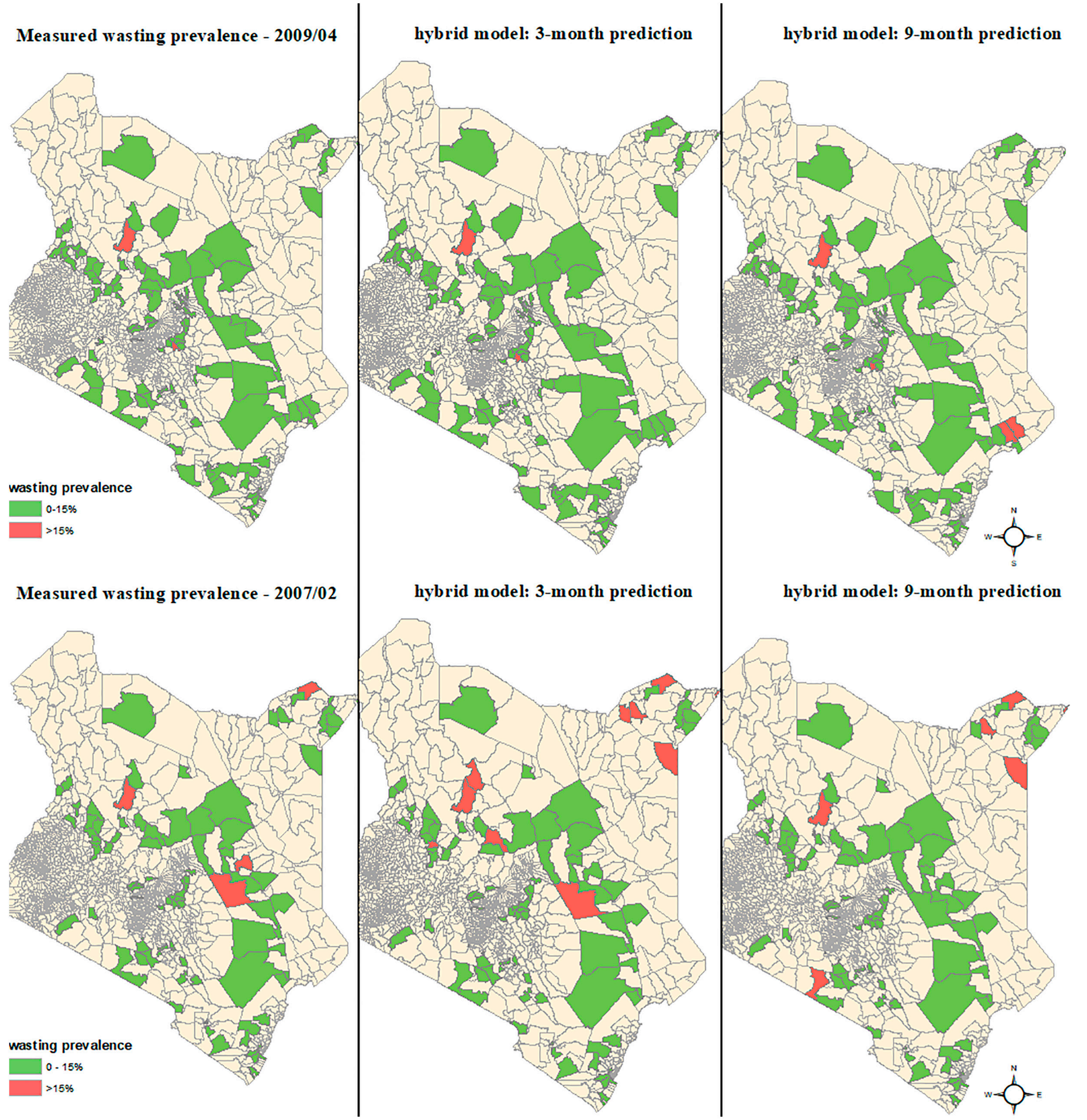
Fig. 3. Gain over time, hybrid model. The figure shows a local polynomial regression of the variable importance in terms of information gain of the hybrid model, for each month considered. The y-axis scale differs among rows to signal the dynamic change of the importance of different variable or variable groups over time. The variables displayed are: GPP (*A*), precipitation (*B*), lags of the wasting prevalence (*C*), static variables (*D*), prices (*E*), and conflict incidence (*F*). Static variables are a sum of: accessibility, population density, land use, elevation, and FEWS Net livelihood zones.

lagged values. The Figure also shows a local polynomial regression of the 1 mo sensitivity of the preferred specification, the hybrid model. Temporal autocorrelation starts decreasing before the data location changes in 2016, reaching its lowest values in 2017. The serial correlation values do not return to values seen in previous periods of the time series until the end of 2018. When temporal autocorrelation is restored, the model's predictive capacity measured in terms of sensitivity goes up again.

A possible explanation for this lack of temporal continuity in the outcome variable is that there were data collection issues during that time period, with the new data collection system still not fully working in an appropriate way. Another plausible explanation is that food emergencies were relieved with a surge in inflows of humanitarian aid, a key variable missing from our dataset. False positives in food security forecasting trained on historic data due to lack of accounting for food aid are commonly reported in the literature (11, 14, 36), and may also explain the drop in specificity that happens during this time period for models that rely on secondary data. Large-scale relief food distribution and cash transfers during the 2016 to 2017 severe drought in Kenya was unprecedented, far exceeding the response during the 2011 drought event (37), which was similarly severe and also recorded in our dataset. This could explain why models relying on secondary data yield a high rate of false positives and why the stationarity of the wasting prevalence predicts and becomes temporarily uncommonly low. This issue would not necessarily be a problem in future early warning tasks done with these models, since high false positive rate due to humanitarian assistance is just an omitted variable bias problem that stems from training on historic data without comparable responses. Another regime shift in humanitarian food response could, however, cause similar degradation in forecast performance in the absence of reliable relief distribution data.

The second factor affecting predictive performance is the increased difficulty of detecting high wasting values when those cases become very rare. High GAM prevalence—food emergencies—are rare events throughout the whole time series. The average number of wards per month with wasting prevalence higher than 15%, the alert threshold, is only 2 out of an average of 130 wards per month for which data are available. Overall, only 371 out of 21,272 ward-month observations in the dataset's time span surpass the 15% intervention threshold. However, food emergencies are more widespread and persistent in some periods than in others.

To see how the model's predictive performance differs between high and low alert periods, we define high alert periods as those with widespread, persistent GAM. Fig. 2*B* shows the cumulative number of months for which there are more than the average of two wards in alert per month. The panel also shows the sensitivity of the 3 mo predictions of the preferred specification (hybrid model) to ease visualization of the model's performance during food emergencies. Performance peaks coincide with high alert periods, when there are two or more wards in alert during at least 2 or 3 consecutive months. This highlights both the difficulty of predicting very rare, outlier-type events, and the predictive capacity of the model for early warning of potentially persistent and relatively widespread food emergencies. It is important to note that even in the high alert periods, the communities in alert correspond to just 2 to 3% of the total, underscoring the model's predictive capacity of relative outliers. High alert periods are


Measured wasting prevalence - 2009/04

hybrid model: 3-month prediction

hybrid model: 9-month prediction

Measured wasting prevalence - 2007/02

hybrid model: 3-month prediction

hybrid model: 9-month prediction

Fig. 4. Example map of measured vs predicted wasting prevalence. Map of measured and predicted (hybrid model) wasting prevalence in two example months of the time series. The maps represent a potential operational outcome of the predictive models.

therefore mostly driven by 3 to 5 wards on alert during those months, never exceeding a total of 8 wards in a given month (of 182 total unique wards in the dataset in the pre-2016 sample and 149 in the post-2016 sample).

A possible solution for increasing the model's capacity for detecting rare events is oversampling those rare cases. We provide results of oversampling of the minority class with ADASYN in the *SI Appendix, Oversampling results*. Overall, oversampling increases predictive performance of very rare cases. But the increased sensitivity comes with a reduction in precision. Close monitoring of predicted prevalence over time at the individual ward level can provide a better approach, as shown in the next section.

Dynamic Fluctuations at the Individual Ward Level. Although the average dynamic results are insightful, the ultimate purpose of such a model is to produce accurate alerts at the individual ward (ADM3) level. Our results clearly show that predictive skill is enhanced by harnessing the autocorrelation of the outcome variable over time. It is therefore natural to ask whether the model can capture trends and predict changes in the evolution of wasting prevalence, or just continuity in the alert or nonalert states.

SI Appendix, Fig. SI5 shows the evolution of the measured and the 1- and 3-mo predicted wasting prevalence over time at the individual ward level for two representative wards (numbers 450 and 1,080) for our preferred specification, the hybrid model. The upper panel of each subfigure contains the measured and predicted wasting prevalence. To aid with the visualization, and to be able to discern trends more clearly, the lower panel shows the (truncated for just past values) Gaussian smoothed values of the same measured and predicted wasting prevalence. The figures also include the upper bound of 95% bootstrapped CI of the predicted values.

These results show how the 1- and 3-mo predictions from the hybrid model encompass the dynamic trends of the measured wasting prevalence. Following the evolution of the upper bound can prove especially helpful when trying to determine whether the wasting prevalence of a particular ward is increasing toward the intervention threshold. Focusing on measuring sensitivity alone over a fixed value undermines the model's ability to help predict when a ward has a high and increasing wasting prevalence, versus a high and decreasing one. Humanitarian response agencies need to separate those two cases, to respond to the first but not necessarily to the latter.

Dynamic Variable Importance. The long time series used in this analysis allows for uncommon exploration of how variable importance changes over time. Fig. 3 shows the information gain of the variables with the highest predictive power included in the preferred specification, the hybrid model.

Important takeaways from the figure are that, first, relative variable importance is dynamic, changing markedly over time. Second, the dynamic variables, not just the static ones, hold predictive power. This is an insightful result that challenges prior claims (20). As can be seen in panel *D*, the static variables (including livelihood zone, remoteness, land use, elevation, and population density) have the highest relative importance at the beginning of the time series, when there are no previous periods on which the model can be trained. This relative importance decreases over time, as the information held in lags of the outcome variable becomes steadily more important. Nevertheless, even when the predictive value of the lags decreases during the low autocorrelation period previously shown, their importance does not peak as much. The importance of other dynamic variables, such as Gross Primary Production (GPP), precipitation, and conflict, increases during this period.

This is an important result only discoverable in a long time series such as this one. Similar forecasting exercises reported in the literature typically report on just single, short time periods, equivalent to any one of our subperiods. Thus, it is not that dynamic variables do not hold predictive power, but rather that one needs higher data collection frequency and an extended period to capture the predictive power of those variables.

Operational Outcomes. Finally, we offer the example of an operationally useful outcome to illustrate how the model could be used as part of a GAM EWS. Although no universal quality thresholds exist for determining the operational usefulness of a forecasting model, we apply two criteria in our case: 1) outperforming TSO and 2) achieving sensitivity and precision above 0.60. The model from the first subperiod satisfies these criteria. Fig. 4 shows a map of the wards in alert according to the measured wasting prevalence and the predictions of the preferred specification (hybrid model) at 3 and 9 mo in advance. The upper panel shows a period with very high sensitivity and precision, where the model can predict with great accuracy the wards with high wasting prevalence. In the lower panel, the predictions are not as accurate, but the wards in alert are spatially proximate to one another, the model highlighting areas where hunger episodes occur. In these cases, contiguously triggered areas could be monitored with 1- or 3-mo in advance forecasts, the ones with the highest sensitivity.

Conclusions

To date, studies using machine learning methods and big, secondary datasets have struggled to generate accurate predictions of anthropometric measures of generalized acute malnutrition, even contemporaneously (i.e., in a nowcasting model), and have failed to forecast future conditions accurately. We hypothesize that the low skill of most such models arises because they are trained on largely cross-sectional or low frequency panel data, meaning that the model cannot learn from the dynamic determinants of malnutrition over time. Using an unusually long monthly data series from an EWS in Kenya, we show that one can forecast U5 child GAM prevalence at a relatively high, operationally useful, spatial resolution (specifically, wards, ADM3) out to 6 mo in advance by retraining models on higher frequency data. Further, we show that disruptions to data collection reduce model performance, that model predictive skill is greatest when one most needs it to work well—as GAM episodes become more widespread and intense – that monitoring of trends, not just predicted levels, at individual ward level can work quite well, and that variable importance varies markedly over time. These findings imply that supervised machine learning methods and big data indeed show promise in assisting in EWS forecasting tasks as a complement to ongoing sentinel site monitoring of child anthropometry. The findings also underscore the role of temporal autocorrelation in generating accurate predictions using supervised machine learning methods in this setting. Finally, our findings emphasize that the intrinsic nonstationarity of anthropometric malnutrition indicators necessitates high frequency data collection to enable regular updating of training data so as to maintain predictive skill in generating EW of child malnutrition.

We employ gradient boosting regression due to its flexibility with various data types, resilient handling of missing values, and overall proven superior performance among ML algorithms for supervised learning with tabular data (38). Given that our outcome variable—the wasting prevalence—contains a substantial proportion of zero values, future research could explore alternative methods better suited for zero-inflated data. One potential approach is a two-stage ensemble model, where the first stage involves a binary classifier distinguishing between zero and nonzero outcomes, and the second stage uses a regression model to predict the magnitude of nonzero values.

Methods

Gradient Boosting Regression for Time-Series Forecasting. We use gradient boosting regression as the base method for our forward-looking predictions. We carry out a walking-forward validation exercise, using sliding windows of 36 mo of data as the training dataset. With those trained models, monthly predictions of the community level wasting prevalence are generated at 1, 3, 6, 9, and 12 into the future. The unit of analysis is ward-month. We use Python's XGBOOST package.

The predictive variables in our dataset are a set of geospatial variables that come mainly from secondary data. To explore the information gain derived from the use of this external data, we define three different gradient boosting regression model specifications that differ in the set of variables they include. Those three models are, 1) a model comprised solely of static and dynamic variables obtained through secondary data (SD), 2) a simple time series model that includes only three lags of the dependent variable (TSO), and 3) a hybrid model that combines the variables included in the TSO and SD models. Both the SD and hybrid models incorporate three lags of the dynamic geospatial variables, including prices and weather measures. Those two models also include 1 y lags of malaria prevalence at the community level, and the seasonal average standardized precipitation and Solar-Induced Chlorophyll Fluorescence-derived gross primary production (GOSIF-GPP) during the long and short rainy seasons (the long rainy season comprises the months from March to June; the short rainy season spans October through December).

Hyperparameter tuning. We carry out hyperparameter selection through the same forecasting exercise we do in our main paper. The sliding windows of historic training data are the same for testing and validation purposes. For validation, however, we just forecast on a randomly selected 20% of the wards in the sample, reserving the other 80% for testing. We do this validation exercise for every month in the time series, that is, training in 36-mo windows of historic data with all wards in them, and predicting in just the selected 20% of wards in future time steps. We select the hyperparameter values with the highest average performance, measured by R^2, across all monthly predictions. With those selected hyperparameters, we retrain the model in each sliding window of historic data, and predict in the remaining 80% of wards in future time steps. This way, we use the same training datasets or sliding windows, methodology, and performance metrics for validation and testing.

We set the total number of trees/iterations to 1,000 and then evaluate and select, 1) maximum tree depth; 2) subsample or fraction of observations that are randomly sampled on each tree; 3) feature downsampling for each tree level; 4) and the learning rate or step-size shrinkage using the hyperparameter tuning strategy described above.

Time-Series Smoothing and CI. We perform time-series Gaussian smoothing to better see trend changes at the individual ward level. We apply a 3-sigma Gaussian kernel to a sliding window of past observations, including the present one, to mimic what could be operationally done in a real early warning application of the models.

To account for uncertainty in individual predictions when following individual ward predictions over time, we generate 95% bootstrapped CI of the forecasts, with 1,000 bootstrap iterations per prediction. We preserve the observation balance of each time-period in the training window when resampling with replacement.

Oversampling Procedure. We perform data augmentation of the minority class through Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) (39). Observations that form part of the minority class are those that identify hunger episodes, i.e. those above the official WHO threshold of 15% wasting community prevalence that triggers an emergency response. We use default number of nearest neighbors and a 20% oversampling ratio (i.e., the number of instances in the minority class has to be at least 20% of the number of observations in the majority class). The oversampling is done for each training 36-mo sliding data window.

Performance Evaluation. To evaluate the OOS performance of our models, we rely on a walking-forward validation exercise. That is, with our already tuned model, we use 36-mo sliding data windows for training, and predict at 1, 3, 6, 9, and 12 mo into the future with the 80% remaining wards not used for hyperparameter tuning of the randomly split data from subsequent time periods. The dataset used for testing has therefore not been used in the training or validation processes at any point, so we can therefore consider the performance metrics obtained through these predictions as an unbiased estimate of the models' OOS performance. We evaluate performance in two different ways: through model fit using R^2, and through classification accuracy using sensitivity, precision, and specificity. We follow the WHO's "wasting prevalence for public health significance" cutoffs (5) and define a wasting prevalence over 15% as the cutoff to determine whether a ward is experiencing a malnutrition crisis in a particular month. Based on that threshold, we calculate the sensitivity, precision, and specificity at every point in time of the data series, for each prediction time horizon.

Data. Mid-upper arm circumference (MUAC) is closely related to other measures of acute malnutrition, such as weight-for-height. Official malnutrition thresholds published by the WHO allow for the identification of GAM through MUAC measurements. The best candidate GAM outcome variable in our study is therefore the prevalence, at the sublocation or ward level, of children under the age of five who present moderate or SAM. These measures are directly targetable by government agencies and other relief institutions seeking to minimize GAM.

As part of its policies to minimize GAM, Kenya's national government, through the NDMA, monitors the subpopulations most vulnerable to drought-related food emergencies, the residents of the country's arid and semiarid lands (ASAL). NDMA tracks conditions in sentinel sites in the 23 ASAL counties among the nation's 47 counties. Prior to a 2010 change to Kenya's administrative structure, the country had been subdivided into 8 provinces, 69 districts, 497 divisions, 1427 locations, and 6612 sublocations. NDMA's sentinel sites were originally chosen in 1999 to be statistically representative of all ASAL livelihood zones and households based on cluster random sampling of a rotating panel of households within 182 different sublocations. MUAC and other data have been consistently collected through field monitors who administer questionnaires and take MUAC measurements of sample households in the sentinel sites. In 2020, NDMA introduced a new system under which mothers were given and trained on the appropriate use of MUAC tapes so that they could take MUAC readings themselves, a move intended to reduce interpersonal contact during the COVID pandemic. NDMA stores the sentinel site data with a custom database, originally the Revised Early Warning Analysis System (40), which in 2016 was absorbed into NDMA's Drought Early Warning System (DEWS).

Since its 2010 constitutional reform, Kenya switched from the provincial administrative system to one based on 47 counties, subdivided into 314 subcounties, which further subdivide into 1,450 wards. In 2010, wards replaced sublocations as the administrative unit within which NDMA samples households. The sentinel sites from the 182 original sublocations mapped to 149 different wards post-2010. We use georeferenced boundaries to match the old sentinel site locations with the new administrative system of counties, subcounties, and wards in creating secondary data series. Sentinel site wards were then reselected in 2016, with just 50% of the original wards included in the post-2016 dataset. Sentinel site locations changed within wards as well. We study the time series created by combining the original dataset that ran from January 2006 to June 2016, with the resampled data that cover July 2016 to December 2020. This not only gives us a longer time series of spatially matched observations, it also permits us to test the impacts of disruptions to time series data collection on forecast performance.

Outcome Variable. We generate our outcome variable, ward-level average prevalence of children under GAM, from monthly MUAC measurements of children under 5 y of age (6 to 59 mo old). Those measurements come from the NDMA sentinel sites and the DEWS database. The sample is representative at the ward level. Following WHO guidelines, wasted children are identified in the sample as those whose MUAC falls below 125 mm. We have complete MUAC information from January 2006 to December 2020, although data collection becomes sparse during 2020 as NDMA transitioned to direct submission of MUAC measures by mothers through DEWS. The unit of analysis is the ward-month level.

Besides matching locations between the pre- and post-2010 administrative units, the only other cleaning performed on the outcome variable data is the elimination of infeasible MUAC values. For the age range considered in the sample, valid observations fall within minimum and maximum thresholds of 80 mm and 250 mm, respectively. A close exploration of the eliminated observations suggests they correspond to data entry errors. Since this step eliminates just 0.26% of total MUAC observations, we do not to attempt to fix them manually or employ other cleaning techniques such as winsorizing; we just eliminate these obvious errors from the sample.

Last, we consider statistically valid monthly prevalences as those generated with at least 50 individual observations. The average number of individual, child MUAC observations per ward is 263 in the pre-2016 dataset, and 141 in the post-2016 one. We examine the eliminated ward-month observations to check whether areas under a food security crisis are consistently undersampled and find no evident patterns of undersampling across space and time emerging from this exploration. The eliminated prevalences are distributed almost equally through all months, although fewer observations per ward exist in early survey years (2006 and 2007).

Predictor Variables. We base our variable selection on indirect or underlying predictors of children's nutritional status [following the framework in (41)], with special focus on the variables relevant in the arid Kenyan context. We also try to incorporate the available variable selection that the NDMA currently uses to monitor food security crises. The final predictors dataset is mainly composed of a set of geospatial variables. To capture the effects of drought in food availability and accessibility we use satellite-imagery products as a proxy for land productivity (GOSIF-GPP, from (42), and weather-related variables [rainfall from (43) and temperature from (44)]. We also control for elevation (45) as a factor influencing the expression of these weather patterns. To control for access to services and more diverse food value chains we incorporate a series of location-specific, indirect measures of development such as accessibility to cities (46), population density (47), and livelihood zone, (48). Crop and rangeland masks (49) are used as measures of the degree of dependence in agricultural or pastoralist livelihoods. To control for fluctuations in food access derived from food price volatility, we retrieve monthly, self-reported, price data on three main food staples of Kenya's arid regions (rice, maize, and beans). Finally, we incorporate monthly, 12-mo running, cumulative incidence of violent conflict events and their associated fatality count (50) and yearly malaria incidence rates in the population (51). Additional details on predictor data sources and variable construction can be found in the *SI Appendix*.

Data, Materials, and Software Availability. Geolocated data files (no individually identifiable data) have been deposited in GitHub (https://github.com/susanaconstenla/Constenla_Villoslada_et_al_2025_PNAS) (52).

ACKNOWLEDGMENTS. This work was funded by the United States Agency for International Development (USAID) under cooperative agreement number 7200AA18CA00014, "Innovations in Feed the Future Monitoring and Evaluation–Harnessing Big Data and Machine Learning to Feed the Future". This work was also undertaken as part of the Food Security Portal project led by the International Food Policy Research Institute, with funding support provided by the European Commission (EC). Neither USAID nor EC had a role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Joshua E. Blumenstock, Nathan Jensen, and a seminar audience at UC-Berkeley for constructive comments on an earlier version.

Author affiliations: [a]School of Information, University of California at Berkeley, Berkeley, CA 94720; [b]Markets, Trade, and Institutions, International Food Policy Research Institute, Washington, DC 20005; [c]Charles H. Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY 14853; [d]Center for Economic Studies, U.S. Census Bureau, Suitland, MD 20746. Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau; [e]National Drought Management Authority, Nairobi 00200, Kenya; and [f]Cornell Jeb E. Brooks School of Public Policy, Cornell University, Ithaca, NY 14853

1. World Food Programme Partnerships and Advocacy Division, *Annual Review, 2023* (World Food Programme Partnerships and Advocacy Division, 2022).
2. Radhini Karunaratne, Jonathan P. Sturgeon, Rajvi Patel, Andrew J. Prendergast, Predictors of inpatient mortality among children hospitalized for severe acute malnutrition: A systematic review and meta-analysis. *Am. J. Clin. Nut.* **112**, 1069–1079 (2020).
3. UNICEF, *The staTe Offood Security and Nutrition in The World 2019* (World Health Organization, 2019).
4. World Health Organization et al, *Guideline: Updates on The Management of Severe Acute Malnutrition in Infants and Children* (World Health Organization, 2013).
5. World Health Organization, *The Management of Nutrition in Major Emergencies* (World Health Organization, 2000).
6. Ananth Balashankar, Lakshminarayanan Subramanian, Samuel P. Fraiberger, Predicting food crises using news streams. *Sci. Adv.* **9**, eabm3449 (2023).
7. B. P. J. Andree *et al.*, *Predicting Food Crises* (The World Bank, 2020) https://ideas.repec.org/p/wbk/wbrwps/9412.html.
8. Erin Coughlan *et al.*, From rain to famine: Assessing the utility of rainfall observations and seasonal forecasts to anticipate food insecurity in east africa. *Food Security* **11**, 57–68 (2019).
9. Daniel Maxwell, Abdullahi Khalif, Peter Hailey, Francesco Checchi, Determining famine: Multi-dimensional analysis for the twenty-first century. *Food Policy* **92**, 101832 (2020).
10. Daniel Maxwell *et al.*, Using the household hunger scale to improve analysis and classification of severe food insecurity in famine-risk conditions: Evidence from three countries. *Food Policy* **118**, 102449 (2023).
11. David Backer, Trey Billing, Validating famine early warning systems network projections of food security in africa, 2009–2020. *Global Food Sec.* **29**, 100510 (2021).
12. P. Krishna Krishnamurthy, Richard J. Choularton, Peter Kareiva, Dealing with uncertainty in famine predictions: How complex events affect food security early warning skill in the greater horn of africa. *Global Food Sec.* **26**, 100374 (2020).
13. Richard J. Choularton, P. Krishna Krishnamurthy, How accurate is food security early warning? evaluation of fews net accuracy in ethiopia *Food Sec.* **11**, 333–344 (2019).
14. S. Arielle *et al.*, Understanding the use of 2015–2016 El Niño forecasts in shaping early humanitarian action in eastern and southern Africa. *Int. J. Disaster Risk Reduct.* **30**, 81–94 (2018).
15. E. C. Lentz, H. Michelson, K. Baylis, Y. Zhou, A data-driven approach improves food insecurity crisis prediction. *World Dev.* **122**, 399–409 (2019).
16. Marshall Burke, Anne Driscoll, David B. Lobell, Stefano Ermon, Using satellite imagery to understand and promote sustainable development. *Science* **371**, eabe8628 (2021).
17. Linden McBride *et al.*, Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning. *Appl. Econ. Perspect. Policy* **44**, 879–892 (2022).
18. Neal Jean *et al.*, Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
19. Andrew Head, Mélanie. Manguin, Nhat Tran, Joshua E. Blumenstock, Can human development be measured with satellite imagery? *ICTD* **17**, 16–19 (2017).
20. Chris Browne *et al.*, Multivariate random forest prediction of poverty and malnutrition prevalence. *PloS ONE* **16**, e0255519 (2021).
21. Christopher Yeh *et al.*, Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nat. Commun.* **11**, 2583 (2020).
22. Christopher Yeh *et al.*, Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. arXiv [Preprint] (2021). https://arxiv.org/abs/2111.04724 (Accessed 1 May 2024).
23. Neeti Pokhriyal and Damien Christophe Jacques, Combiningdisparatedatasourcesforimprovedpovertypredictionandmapping. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E9783–E9792 (2017).
24. Francesco Checchi *et al.*, Can we predict the burden of acute malnutrition in crisis-affected countries? findings from somalia and south sudan. *BMC Nut.* **8**, 1–20 (2022).
25. Andrew G. Mude, Christopher B. Barrett, John G. McPeak, Robert Kaitho, Patti Kristjanson, Empirical forecasting of slow-onset disasters for improved emergency response: An application to kenya's arid north. *Food Policy* **34**, 329–339 (2009).
26. Shahrzad Gholami *et al.*, Food security analysis and forecasting: A machine learning case study in southern malawi. *Data & Policy* **4**, e33 (2022).
27. Pietro Foini, Michele Tizzoni, Giulia Martini, Daniela Paolotti, Elisa Omodei, On the forecastability of food insecurity. *Sci. Rep.* **13**, 2793 (2023).
28. Giulia Martini *et al.*, Machine learning can guide food security efforts when primary data are not available. *Nat.Food* **3**, 716–728 (2022).
29. David Backer, Trey Billing, Forecasting the prevalence of child acute malnutrition using environmental and conflict conditions as leading indicators. *World Dev.* **176**, 106484 (2024).
30. B. Christopher Barrett, Measuring food insecurity. *Science* **327**, 825–828 (2010).
31. Derek Headey, Christopher B. Barrett, Measuring development resilience in the world’ spoorestcountries. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11423–11425 (2015).
32. Erwin Knippenberg, Nathaniel Jensen, Mark Constas, Quantifying household resilience with high frequency data: Temporal dynamics and methodological options. *World Dev.* **121**, 1–15 (2019).
33. Giulia Martini *et al.*, Nowcasting food insecurity on a global scale. medRxiv [Preprint] (2021). https://doi.org/10.1101/2021.06.23.21259419 (Accessed 1 May 2024).
34. Erwin Knippenberg, Hope C. Michelson, Erin C. Lentz, Tess Lallemant, *Early Warning; Better Data, Better Algorithms Improved Food Insecurity Predictions Using Machine Learning and High-frequency Data in Malawi* (unpublished manuscript, 2020).
35. V. Flovik, How(not) to use machine learning for time series forecasting: Avoiding the pitfalls. *Medium*, DD MMMM YYYY. https://medium.com/data-science/how-not-to-use-machine-learning-for-time-series-forecasting-avoiding-the-pitfalls-19f9d7adf424. Accessed 16 May 2025.
36. Yujun Zhou, Erin Lentz, Hope Michelson, Chungmann Kim, Kathy Baylis, Machine learning for food security: Principles for transparency and usability. *Appl. Econ. Perspect. Policy* **44**, 893–910 (2022).
37. C. Funk *et al.*, Contrasting Kenyan resilience to drought: 2011 and 2017, USAID Special Report, 20 pages.
38. Duncan McElfresh *et al.*, When do neural nets outperform boosted trees on tabular data? arXiv [Preprint] (2024). https://arxiv.org/abs/2305.02997 (Accessed 1 May 2024).

39. Haibo He, Yang Bai, Edwardo A. Garcia, Shutao Li, "Adasyn:Adaptive synthetic sampling approach for imbalanced learning" in *2008 IEEE International Joint Conference on Neural Networks* (IEEE World Congress on Computational Intelligence, 2008), pp. 1322–1328.
40. Drought Preparedness Intervention and Recovery Project User's Guide, Version 1.0. (Tech. Rep., Revised Early Warning Analysis System, Nairobi, (1999).
41. Robert E. Black *et al.*, Maternal and child undernutrition:global and regional exposures and health consequences. *Lancet* **371**, 243–260, (2008).
42. Xing Li, Jingfeng Xiao, A global, 0.05-degree product of solar-induced chlorophyll fluorescence derived from oco-2, modis, and reanalysis data. *Remote Sens.* **11**, 517 (2019).
43. Chris Funk *et al.*, The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data* **2**, 150066 (2015).
44. J. Muñoz Sabater *et al.*, Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383 (2021).
45. A. Jarvis, H. I. Reuter, A. Nelson, Ed. Guevara, *Hole-Filled SRTM for The Globe version4*. (CGIAR-CSI SRTM, 2008).
46. D. J. Weiss *et al.*, A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553**, 333–336 (2018).
47. CIESIN, *Gridded Population of The World, version 4* (Center for International Earth Science Information Network CIESIN, Columbia University, 2018) **vol. 11**, p. 20240110.
48. OCHA, *Famine Early Warning Systems Network* (Kenya livelihood zones, 2011). https://fews.net/east-africa/kenya/livelihood-zone-map/march-2011.
49. A. Pérez-Hoyos, *Global Cropand Rangeland Masks* (European Commission, 2018). http://data.europa.eu/89h/jrc-10112-10005.
50. Clionadh Raleigh, r. Linke, H. Hegre, J. Karlsen, Introducing ACLED: An armed conflict location and event dataset. *J. Peace Res.* **47**, 651–660 (2010).
51. Daniel J. Weiss *et al.*, Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: A spatial and temporal modelling study. *Lancet* **394**, 322–331 (2019).
52. S. Constenla-Villoslada, Constenla_Villoslada_et_al_2025_*PNAS*. Github. https://github.com/susanaconstenla/Constenla_Villoslada_et_al_2025_PNAS. Deposited 17 May 2025.