Supplemental Materials for *Monitoring and Projecting Global Hunger*

S1   Overview of the FIES

We use raw microdata released by the FAO from 75 countries from different world regions and income categories (See Fig. S1).

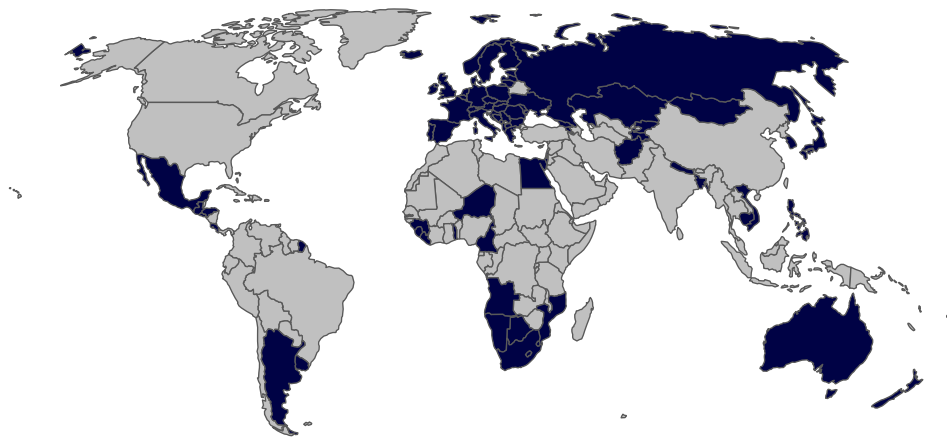### Countries That Have FIES Microdata

Figure S1: Countries used in our model, shown in blue.

The microdata includes information about the probability that an individual is over the threshold for moderate and severe food insecurity, calculated by the FAO on the basis of the responses to the 8 FIES questions (See List 8). The FIES is calculated using a Rasch model, which was developed by the psychometrics literature. The Rasch model assumes that each individual and their responses to the FIES questions can be placed on a one-dimensional scale of food insecurity, and that the log odds of a respondent answering affirmatively to one of the FIES questions is a linear function of the difference between the severity of the food insecurity experienced by the individual and the severity of the item put forward in the corresponding question.

The severity of each item and the respondent's level of food insecurity can be estimated making use of maximum likelihood methods. The assumptions of the Rasch model hold up well for the data at hand, as evidenced by model fit diagnostics [1].

After separately estimating the Rasch model for each country, FAO develops a global reference scale for the severity of each of the eight questions by iteratively harmonizing the severity values in each country. These global reference points are calibrated against other surveys that ask questions similar to those in the FIES, such as the HFSSM and the ELCSA.

Based on the global reference scale, two thresholds are set based on questions 1 and 8, corresponding to moderate and severe food insecurity, respectively.

1. During the last 12 months, was there a time when you were worried you would not have enough food to eat because of a lack of money or other resources?

2. Still thinking about the last 12 months, was there a time when you were unable to eat healthy and nutritious food because of a lack of money or other resources?

3. Was there a time when you ate only a few kinds of foods because of a lack of money or other resources?

4. Was there a time when you had to skip a meal because there was not enough money or other resources to get food?

5. Still thinking about the last 12 months, was there a time when you ate less than you thought you should because of a lack of money or other resources?

6. Was there a time when your household ran out of food because of a lack of money or other resources?

7. Was there a time when you were hungry but did not eat because there was not enough money or other resources for food?

8. During the last 12 months, was there a time when you went without eating for a whole day because of a lack of money or other resources?

S2 Predicting Covariates Into the Future

S2.1 Annualized Rate of Change

For many of our covariates, to extrapolate recent trends to the future, we draw on historically annualized rates of change, which are used to project the future change in the variable. We use this method to estimate future rates of stunting, wasting, and malaria mortality, as well as to model future subnational shares of GDP and population.

For each subnational area, a, the rate of change (ROC) is calculated between each pair of adjacent years, y, based on a value, p, such that $0 < p < 1$, available for each year.

$$\text{ROC}_{a,y} = \text{logit}\left(\frac{p_{a,y}}{p_{a,y-1}}\right) \tag{1}$$

Each ROC is weighted so as to give proportionally more weight to recent years. For a dataset with observations for 2010 to 2019, that would be:

$$W_y = \frac{y - 2010}{\sum_{y=2010}^{2019} y - 2010} \tag{2}$$

The annualized rate of change (AROC) is calculated as:

$$\text{AROC}_a = \text{logit}\left(\sum_{y=2010}^{2019} W_y \times \text{AROC}_{a,y}\right) \tag{3}$$

Finally, the projections, Proj, for each year up to 2030 are given by:

$$\text{Proj}_{a,y} = \text{logit}^{-1}(\text{logit}(p_{a,2019}) + AROC_a \times (y - 2019)) \tag{4}$$

For cases where we model shares with a summing constraint, such as the regional shares of a variable measured at the national level, we rescale the shares to ensure that the constraint holds.

S2.2 Population

Although population is not a covariate in our random forest model, it is an important determinant of other covariates such as GDP per capita. Future estimates of population are also necessary to convert modeled future rates of food insecurity into total headcounts of food insecurity.

We project population by combining historical subnational data from the Subnational Development Database [2] with national level population estimates for the 21st century [3]. To account for within-country trends and changes in population distribution, we use the AROC method outlined in equations 1 - 4 to project each subnational areas share of national population totals. We then disaggregate the national totals given by KC et al. to estimate future subnational populations.
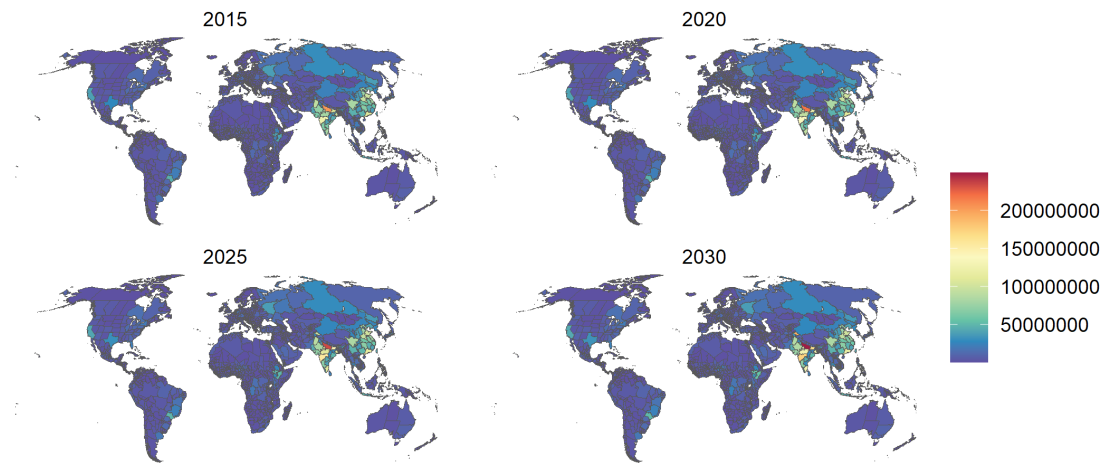

Figure S2: Population

S2.3   Urbanization Rates

We draw our historical and future estimates of urbanization entirely from the spatially explicit scenarios of urban population by Jones and O'Neill [4], which provide a series of urbanization rate projections which are consistent with the Shared Socioeconomic Pathways. As with our other covariates, we use projections consistent with SSP2, the middle-of-the-road scenario.
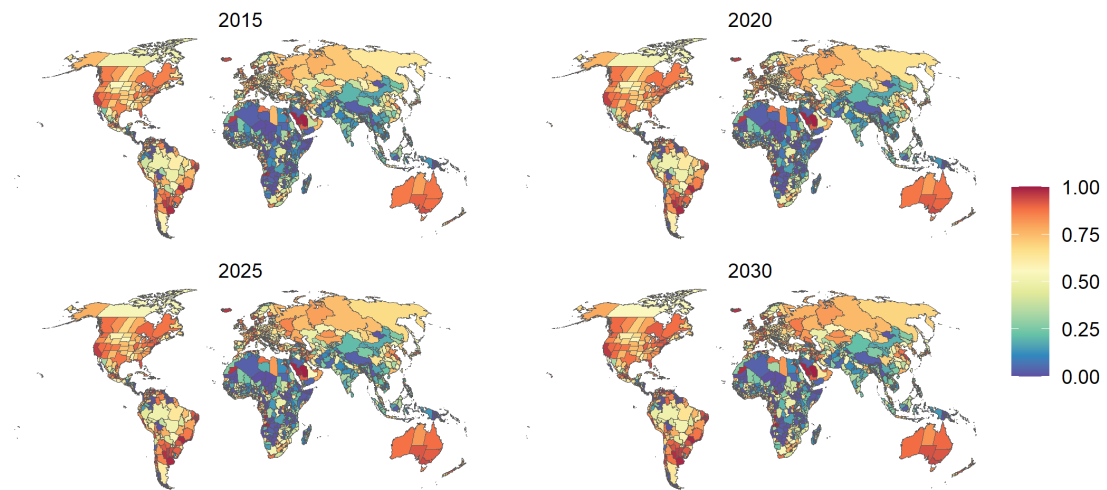

Figure S3: Percentage of people living in urban areas.

S2.4   Wasting

To estimate the prevalence of wasting for each administrative area in our dataset, we use data from the Local Burden of Disease project. For the years 2010 to 2017, we simply use the mean rate of wasting in each administrative area from the dataset. Higher-income countries that are not included in the dataset are assumed to have zero wasting. We then project wasting for the years 2018-2030 using the AROC method, which the Local Burden of Disease group similarly use to estimate wasting for the year 2025 [5].

To account for the effects of the coronavirus pandemic on global rates of wasting, we assume that long-term trends in rates of wasting would hold steady, after they increase globally by 14.3% in 2020, based on estimates published in *The Lancet* [6]. We model this impact as uniform across all countries where wasting occurs. We project the rate of wasting in each administrative area in 2021 as being the average of (a) the rate of change in 2020 accounting for the pandemic shock and (b) the predicted rate of change for 2022 obtained ignoring the pandemic.
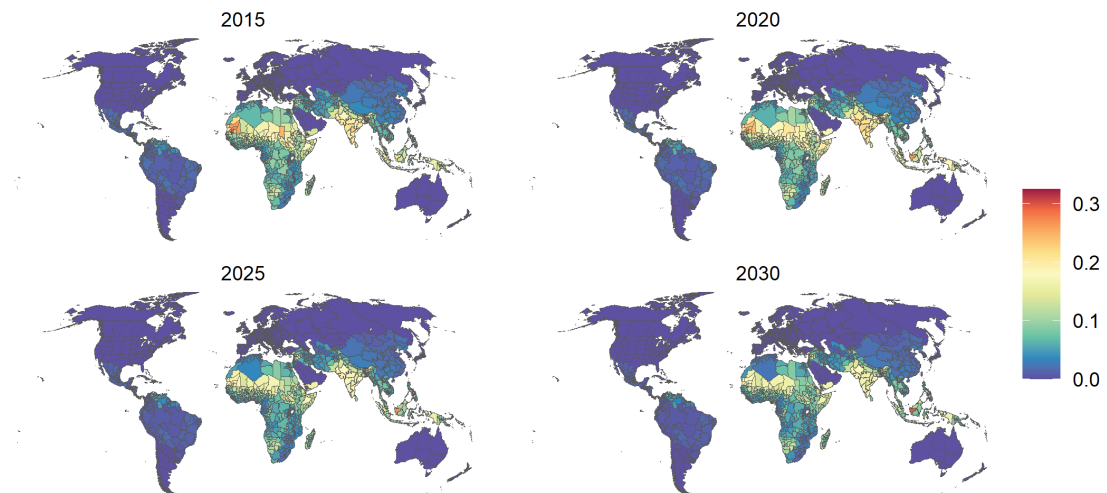

Figure S4: Prevalence of wasting.

S2.5 Stunting

We model stunting, sourced from the Local Burden of Disease dataset, using the same methodology employed for wasting. We also assume a 14.3% increase in prevalence for the year 2020 [6], as is done for wasting, although stunting is a long-term consequence of malnutrition and is not likely to be observable in a population with the same lags as wasting would be. We use stunting in our modeling framework as a proxy for chronic conditions of hunger, which are probably exacerbated by the coronavirus pandemic.
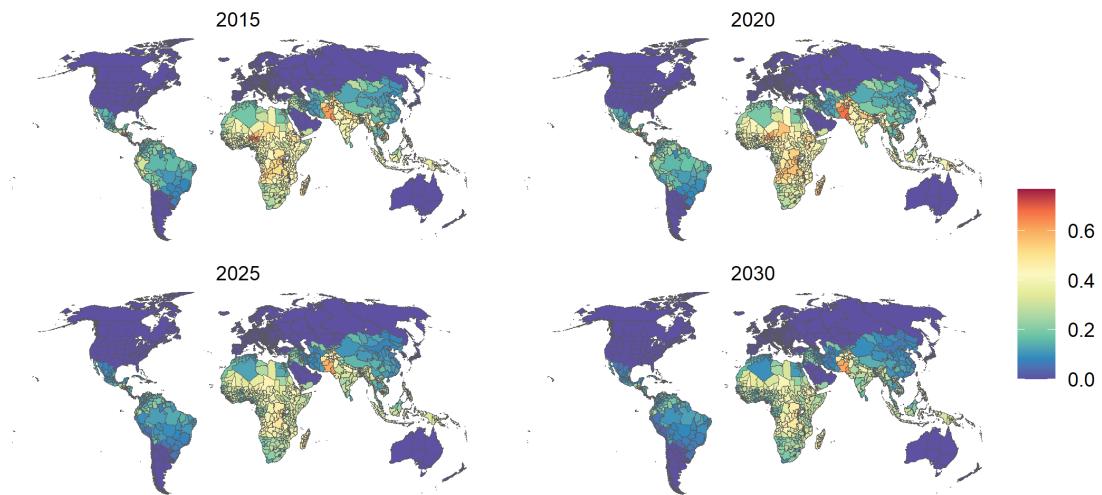

Figure S5: Prevalence of stunting.

S2.6 GDP Per Capita

We estimate global subnational GDP per capita using historic subnational estimates of Gross National Income (GNI) from the Subnational Human Development Database [2] and forecasts of national GDP within the SSP framework from Dellink et al. [7]. We first use data from the World Bank to estimate the mean country-level ratio of GNI to GDP and adjust the subnational estimates of GNI according to that country-specific ratio. We then calculate each subnational area's share of country GDP, and then use the AROC method to estimate each subnational area's share of total country GDP. Based on these trajectories in each subnational areas share of total country GDP, we estimate what each subnational area's share of GDP will be for each year from 2019 to 2030. We divide that share of GDP by the population of the administrative area, which we had previously estimated. Thus, national GDP totals are consistent with the projections of Dellink et al., but are disaggregated according to the historic patters of GDP share in the Subnational Human Development Database.

Finally, we used estimates of country-level changes in GDP growth for the years 2020 and 2021 to account for the impact of the coronavirus [8]. We then model GDP as resuming its previous growth trajectory based on previously-estimated growth rates from 2022-2030, but growing from an estimate for 2021 that includes the coronavirus shock.

We measure global subnational GDP per capita using historic subnational estimates of Gross National Income (GNI) from the Subnational Human Development Database [2] and forecasts of national GDP within the SSP framework form Dellink et al. [7]. We source data on the mean country-level ratio of GNI to GDP from the World Bank and adjust the subnational estimates of GNI according to the ratio for each country in our sample. We calculate the share of national GDP for each subnational region, and use the AROC method to project each subnational area's share of total country GDP up to 2030. National GDP totals are consistent with the projections of Dellink et al., but are disaggregated according to the historic patterns of subnational GDP shares sourced from the Subnational Human Development Database. We use nowcasts and forecasts of country-level changes in GDP growth for the years 2020 and 2021 to account for the impact of the coronavirus [8]. After the shock created by the COVID pandemic, GDP is assumed to resuming its previous growth trajectory for the period 2022-2030.
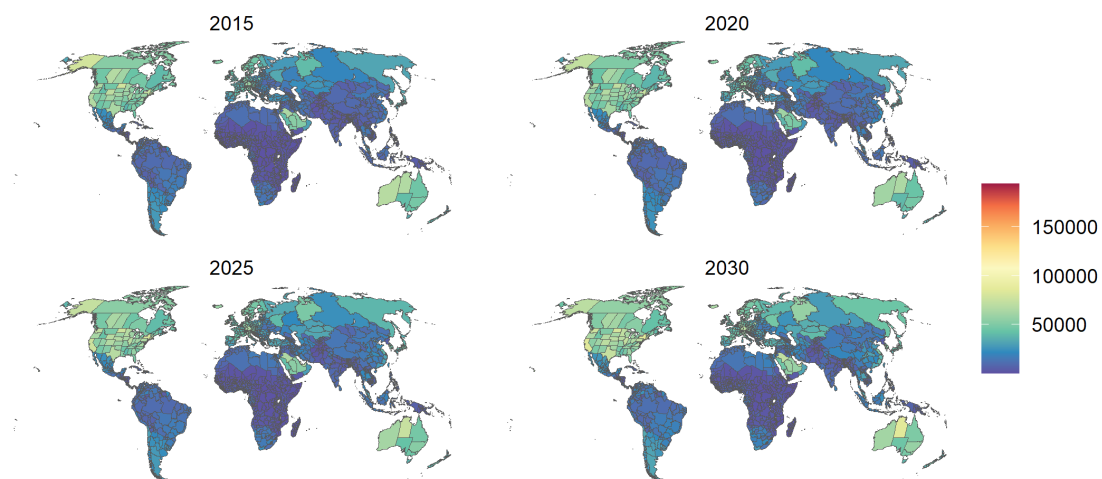
Figure S6: GDP per capita.

S2.7 Mean Years of Schooling

We use data on global subnational mean years of schooling from the Subnational Human Development Database [2] and forecasts of national mean years of schooling from KC et al. [3]. Using the subnational data, we calculate the difference in years of schooling form the nation mean each subnational administrative area, and use this value to disaggregate the future projections to a subnational level, keeping these differences constant over the projection horizon.
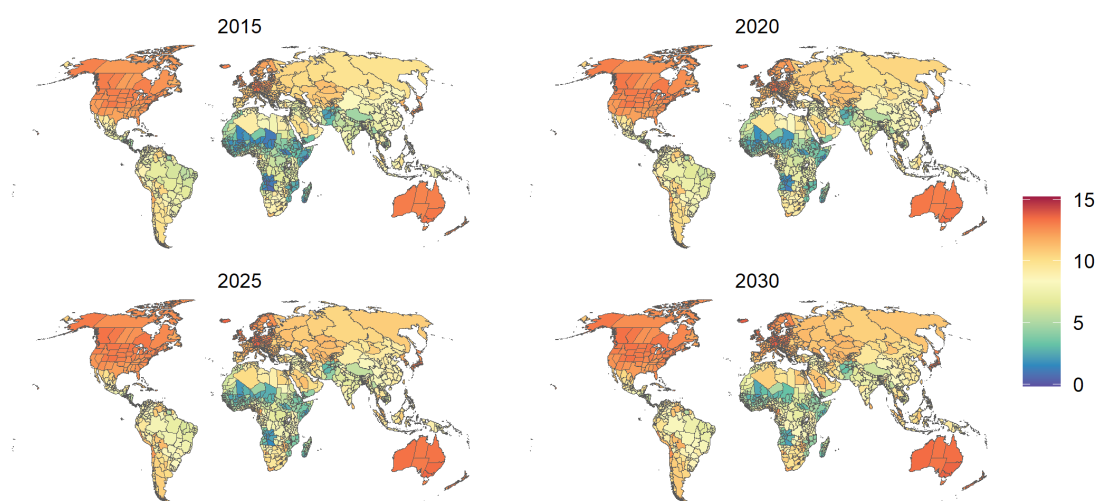
Figure S7: Mean years of schooling.

S2.8   Gini Coefficient

We use data on historical levels and projections of national income inequality from Rao et al.
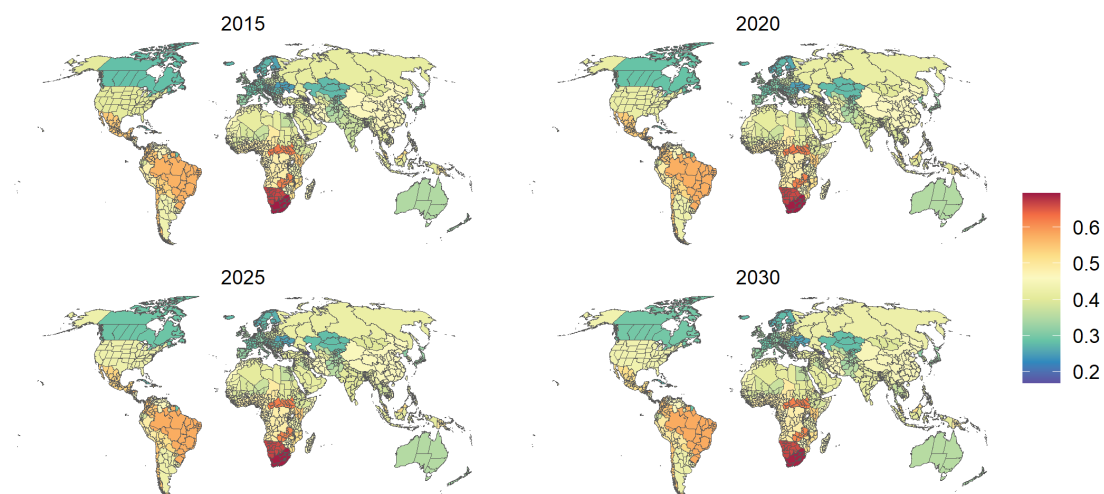[9]. These estimates are not disaggregated to a subnational level.

Figure S8: Gini Coefficient

S2.9 Poverty Headcount Index

We used historical and projected data on the Poverty Headcount Index (using the standard threshold of $1.90 a day) from the World Poverty Clock by the World Data Lab, which in turn uses Povcal data sourced from the World Bank [10]. These data are the version corresponding to the update in summer 2020 and include the expected impact of the COVID pandemic on global poverty dynamics. Using parameterized Lorenz curves in combination with mean income/consumption and population projections, poverty rates are projected.
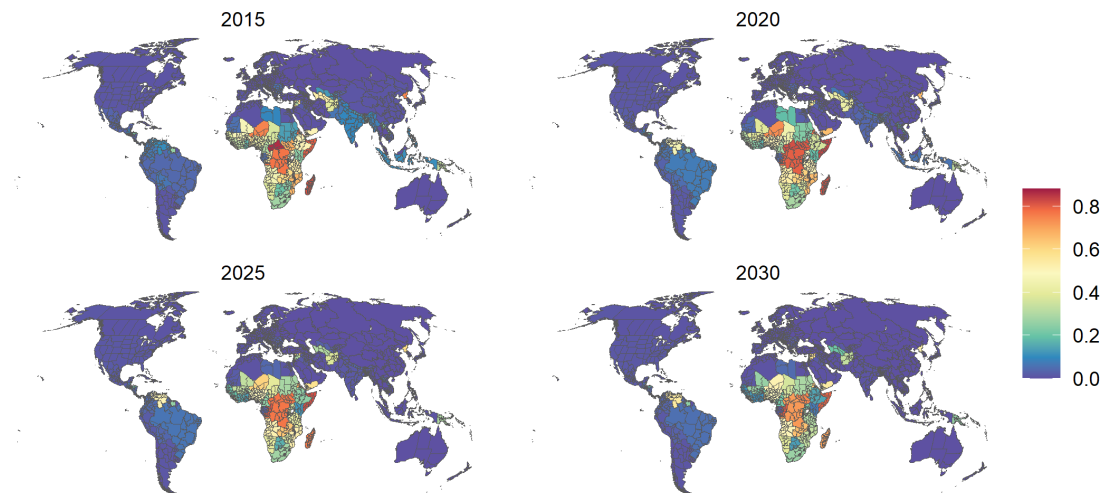

Figure S9: Poverty Headcount Index

S2.10 Water Scarcity

We use data for the water scarcity index (WSI) developed by [11]. The index represents the ratio of the decadal averages in water demand to water supply, based on population projections consistent with SSP2 and water availability projections from global hydrological and general circulation models. Based on this index, we estimate the share of people living in areas with less than 1000m^3 of water available per capita per year. The WSI has been used already to build the water scarcity clock.
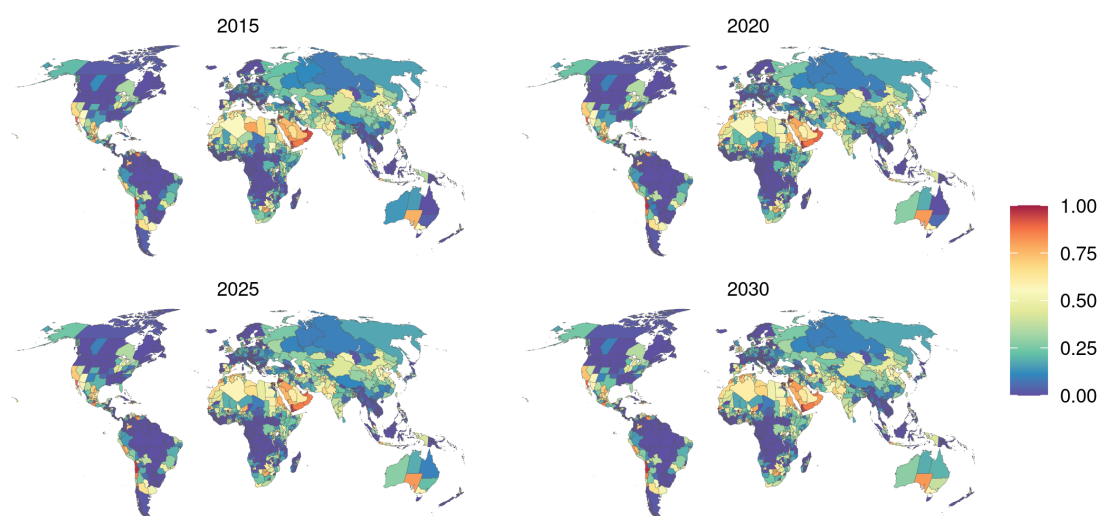

Figure S10: Water Scarcity Share

S2.11 Mean Annual Precipitation and Mean Temperature

For mean annual precipitation and temperature, we use data for the years 2010 to 2019 from the TerraClimate dataset [12], aggregated to each subnational area. For data for the years 2020-2030, we use projections from models from the Inter-Sectoral Model Intercomparison Project [13]. Specifically, we use the mean of an ensemble of bias-corrected projections under representative concentration pathway (RCP) 6.0.
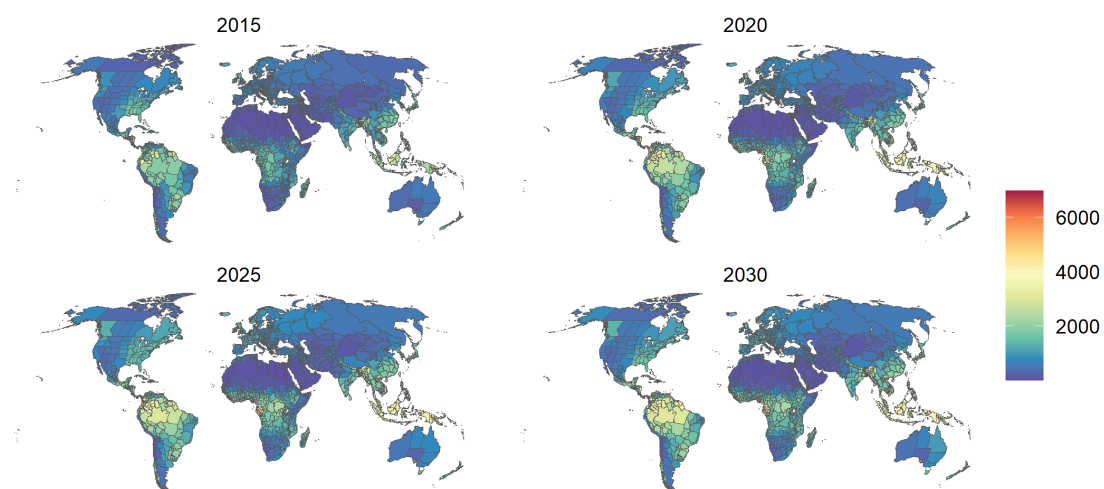
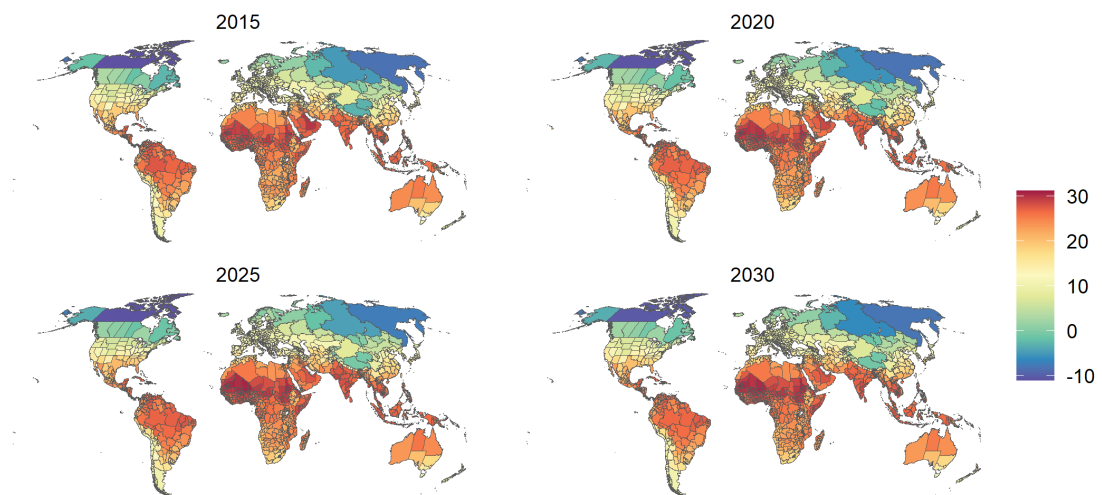

Figure S11: Mean Annual Precipitation

Figure S12: Mean Temperature (Celsius)

S2.12 Topographic Ruggedness

As a measure of accessibility, our models include the mean topographic ruggedness of each subnational area as a covariate. We obtain the variable using a gridded dataset of elevation from the USGS [14] and calculate the index at each grid cell using the methodology from Riley et al. [**Riley1999**]. We then aggregate the values for all the grid cells within each administrative area. The variable is constant in the estimation and projection period.
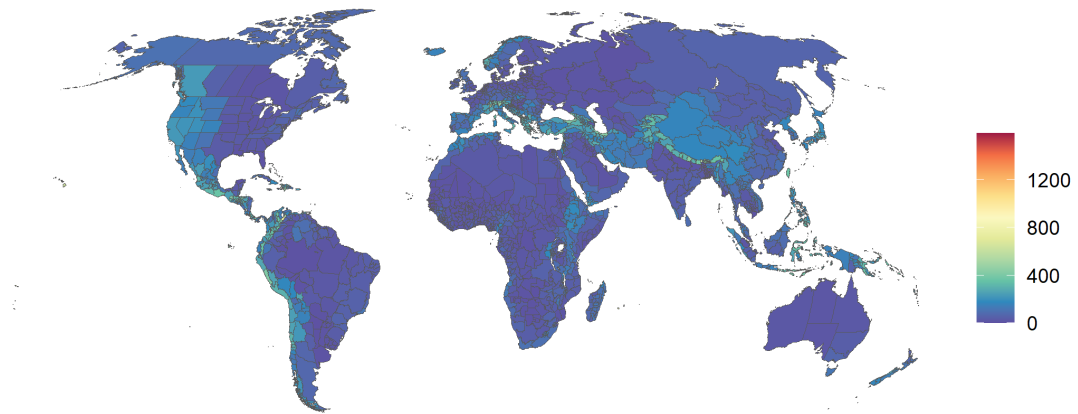

Figure S13: Topographic ruggedness

S2.13   Malaria (*P. falciparum*) Mortality Rate

We used data from Weiss et al. to estimate the mortality (deaths per 100,000 population per year) due to Malaria [15], and aggregate the information at the level of subnational administrative area. Projections of malaria mortality are based on the AROC method.
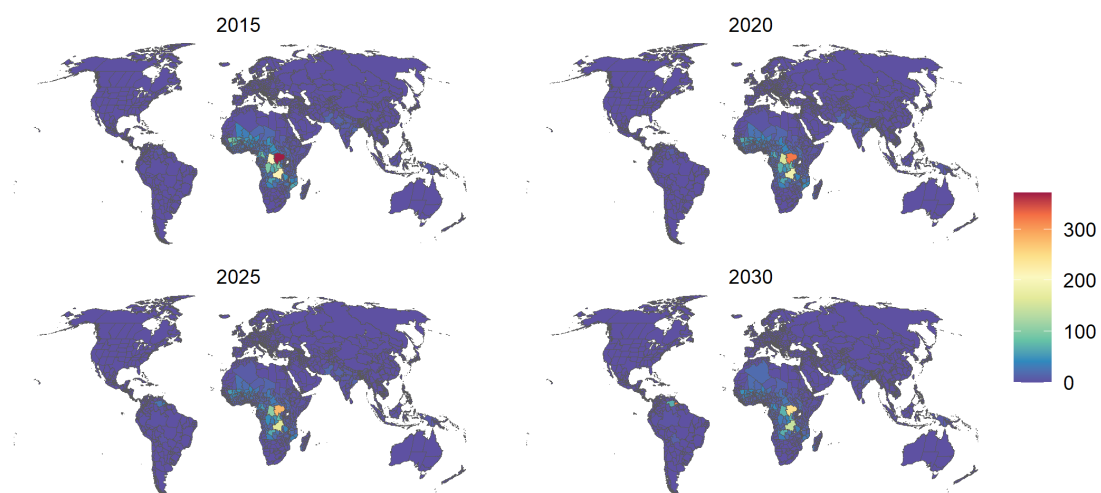

Figure S14: Rate of mortality due to *P. falciparum* Malaria

S3 Description, Implementation and Validation of the Random Forest Regression Model

The random forest regression is a machine learning method based on the creation of a large number of decision trees. For each of the decision trees in a random forest model, observations are selected at random with replacement, a method known as bootstrap aggregating, or bagging, and features are also selected at random. For the implementation we use the R-package `randomForestSRC` by Ishwaran and Kogalur [16].

Because the outcome variables in our models, the rate of moderate or severe food insecurity, are a bounded outcome, we first used a logistic transformation to extend the range from $[0, 1]$ to $(-\infty, \infty)$. Then, after estimating the model, we used the inverse logit to convert the unbounded model predictions to a rate between 0 and 1.

To determine the best hyperparameters to model both moderate ands severe food insecurity, we use leave-one-out cross validation at the country level. This involves running a separate model for each country, where we train a model on the remaining data from all of the other countries, and use that model to predict values for the country that was left out. Then, across the predicted values for all countries, we evaluate the mean average error (MAE) and R^2 of the differences between observed rates of food insecurity and predictions from the cross validation models.

We evaluate models across a grid of hyperparameters, with combinations of: between 1 to 13 variables randomly selected as candidates for splitting a node (`mtry`), between 1 and 20 terminal node sizes (`nodesize`), and with a maximum tree depth (`depth`) between 1 to 13 nodes, as well as without any maximum tree depth. For each model in cross validation, we used 1000 trees (`ntree`). Due to the large size of the hyperparameter space, we sampled sets of hyperparameters randomly, eventually sampling about 50% of all combinations of hyperparamters. We found that the optimal model for moderate food insecurity had 16 average terminal nodes, 2 variables selected as candidates for splitting a node, and no maximum tree depth; the optimal model for severe food insecurity also had 2 variables selected as candidates for splitting a node and no maximum tree depth, but with only 14 average terminal.

After fitting models with cross validation, we evaluated the models performance and graphed a comparison between the observed and predicted values based on instances where individual countries were left out (See Fig. S15), as well as for a model trained on the entire dataset, which can only be evaluated on in-sample data (See Fig. S16). Finally, Figure S17 shows the model error converging and the number of trees increases.
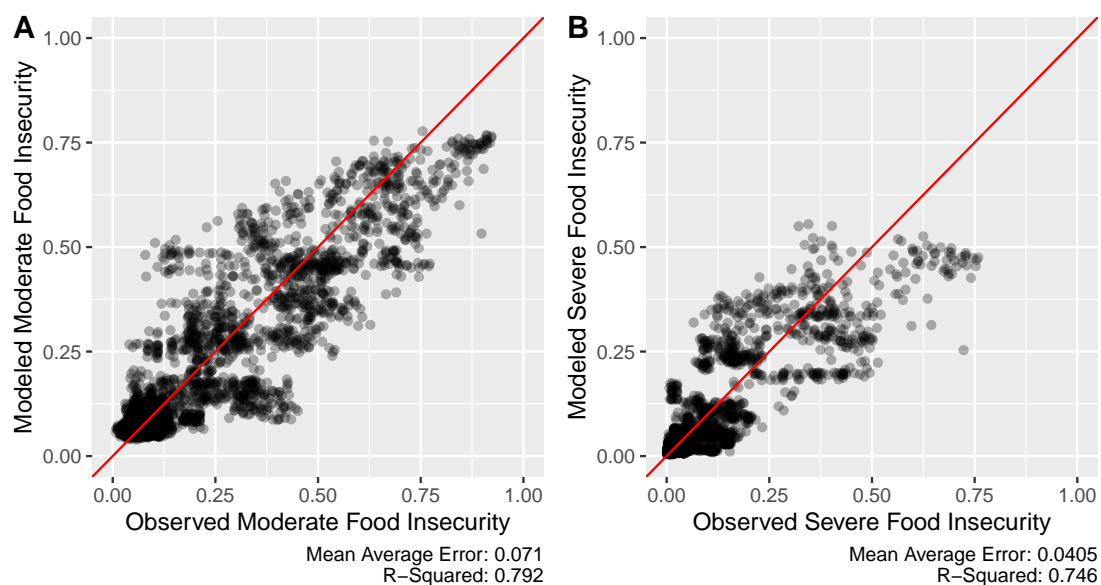
Figure S15: Out-of-sample fit of the random forest regression model, where predictions for each country are based on a model that did not include that country in the training data. Panel (A) shows the model for moderate food insecurity and panel (B) shows the model for severe food insecurity.
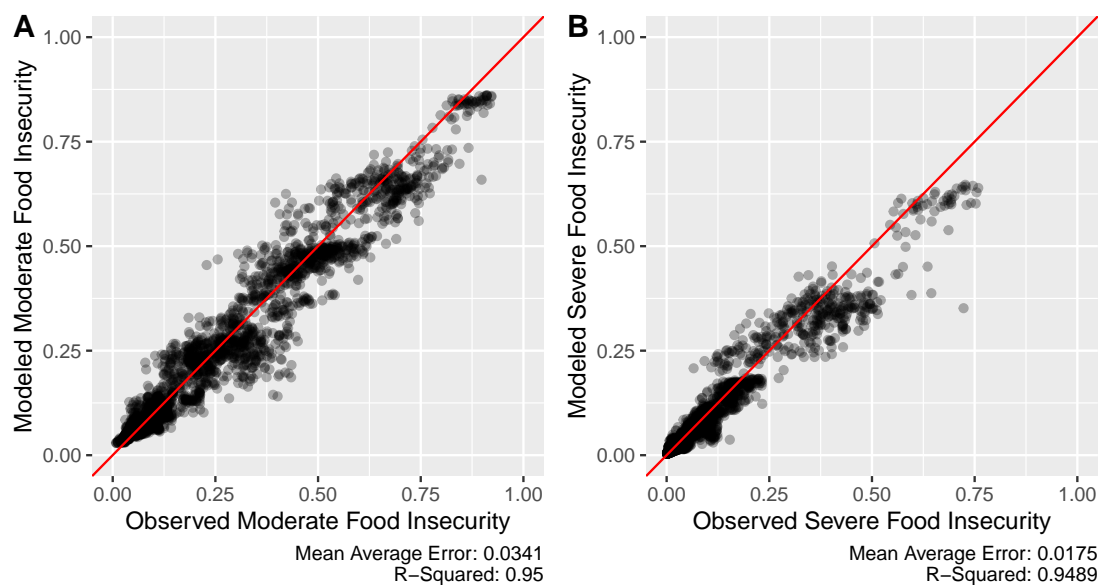

Figure S16: In-sample fit of the random forest regression model, where one model was trained on the entire dataset. Panel (A) shows the model for moderate food insecurity and panel (B) shows the model for severe food insecurity.
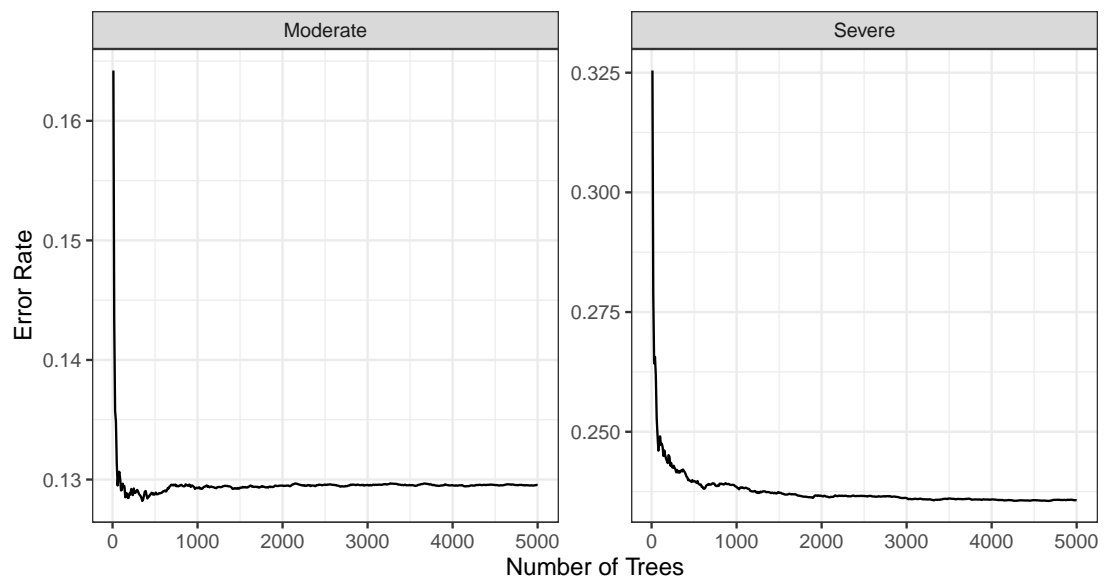
Figure S17: Error rate of the random forest regression model as the number of trees in the model increases. Panel (A) shows the model for moderate food insecurity, and panel (B) shows the model for severe food insecurity.