

Supplemental Info for *A Global Picture of Hunger*

S1 Overview of the FIES

We used raw microdata released by the FAO from 75 countries that vary by world region and income level (See Fig. S1).

Countries That Have FIES Microdata

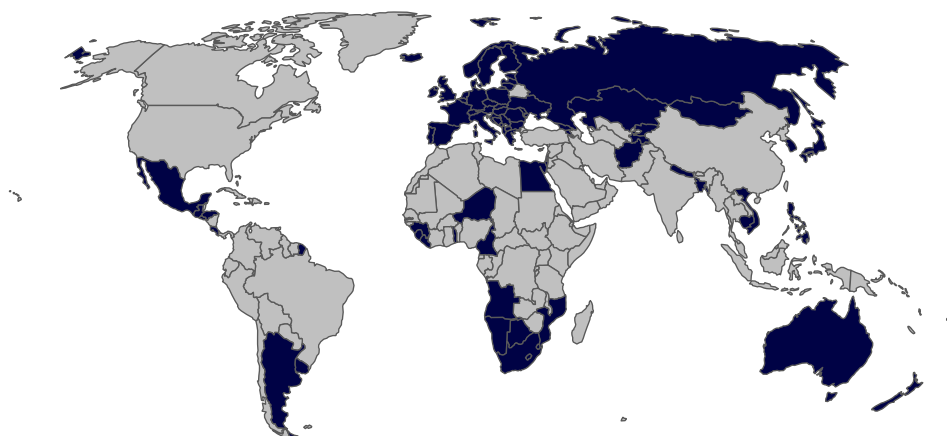


Figure S1: Countries that had microdata on the FIES and were used in our model, shown in blue.

This microdata comes with the probability that an individual is over the threshold for moderate and severe food insecurity already calculated by the FAO, based on their responses to the 8 FIES questions (See List 8). Thus, we did not perform those calculations ourselves. Nevertheless, we give here an overview of the procedure. For more detail see Cafiero (2018) in *Measurement* [1].

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 r answering yes to one of the FIES questions i is a linear function of the difference between the severity of the food insecurity experienced by r and the severity of item i .

The severity of each item and each respondents level of food insecurity can be estimated with Maximum Likelihood. After fitting the model to each country, the FAO found that the assumptions of the Rasch model held up well, as evidenced by model fit diagnostics.

After separately estimating the Rasch model for each country, the FAO then developed a global reference scale for the severity of each of the 8 questions, by iteratively harmonizing the severity values in each country. Then, these global reference points were calibrated against other surveys that had asked questions similar to 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: whether the respondent was worried they would not have enough food to eat because of a lack of money or other resources, and whether there was a time when they went without eating for a whole day because of a lack of money or other resources.

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, it is necessary to draw on annualized rates of change, which are then used to anticipate the rate of future change. This method can be used for any variable that is a fraction or a rate. 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. Thus, we give an overview of the methodology here.

First, 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.

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

Then, each ROC is weighted to give more weight to recent years, such that weight W at year y is. For a dataset with observations for 2010 to 2017, that would be:

$$W_y = \frac{y - 2010}{\sum_{y=2010}^{2017} y - 2010} \quad (2)$$

Then, the annualized rate of change (AROC) is calculated as:

$$\text{AROC}_a = \text{logit} \left(\sum_{y=2010}^{2017} W_y \times \text{ROC}_{a,y} \right) \quad (3)$$

Finally, the projections, Proj, for each year from 2018 to 2030 are given as:

$$\text{Proj}_{a,y} = \text{logit}^{-1}(\text{logit}(p_{a,2017}) + \text{AROC}_a \times (y - 2017)) \quad (4)$$

For cases where we are modeling shares of a whole, for example when modeling subnational shares of population or GDP, we then re-scale the shares to ensure they total to 1.

S2.2 Population

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

We do this by combining historic 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.

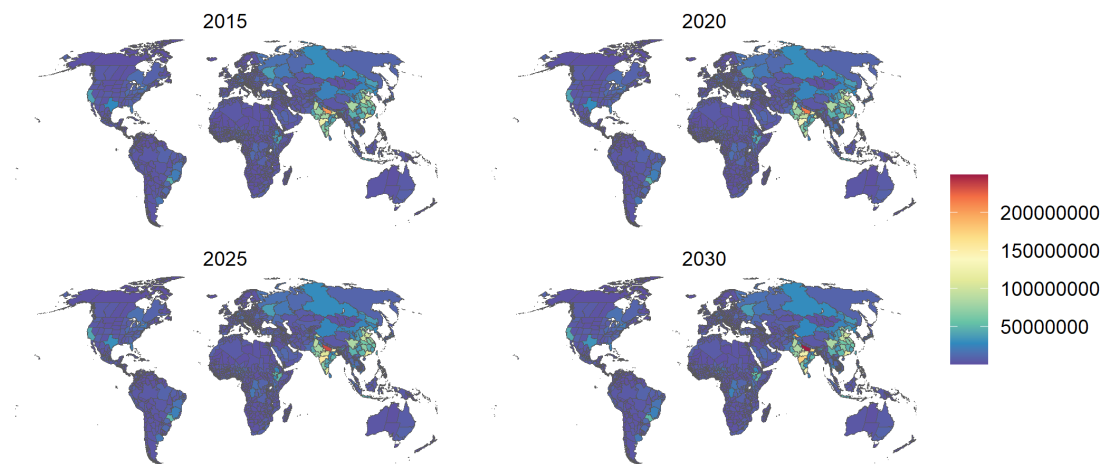


Figure S2: Population

S2.3 Urban Percentage

We drew our historic and future estimates of urbanization entirely from Jones and O'Neill, published in *Environmental Research Letters* [4], who modeled spatially explicit scenarios of urbanization consistent with the Shared Socioeconomic Pathways. As with our other covariates, we used estimates for SSP2, the middle-of-the-road scenario. Because definitions of “urban” and “rural” vary significantly across datasets, using one dataset that is entirely consistent within itself was an important consideration.

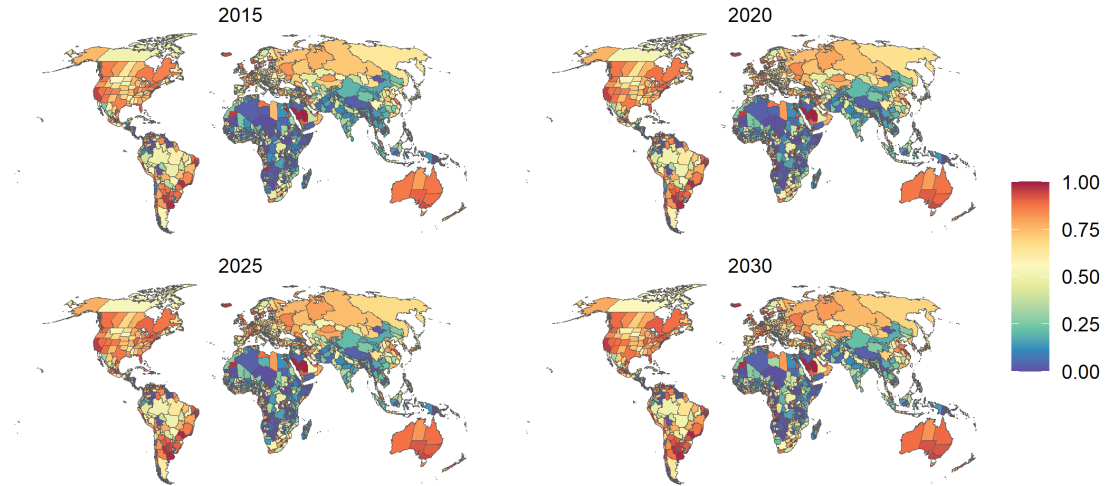


Figure S3: Percentage of people living in urban areas.

S2.4 Wasting

To estimate the prevalence of wasting for each administrative area in our dataset, we used data from the Local Burden of Disease project published in *Nature* [5]. For the years 2010 to 2017, we simply took the mean rate of wasting in each administrative area from the dataset. Higher-income countries that were not included in the dataset were modeled as having a rate of 0 wasting. We then modeled wasting for the years 2018-2030 using the AROC method, which the Local Burden of Disease group similarly used to estimate wasting for the year 2025 [5].

To account for the effects of the novel coronavirus on global rates of wasting, we assumed that long-term trends and rates of wasting would hold steady, but that 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. Then we model the rate of wasting in each administrative area in 2021 as being the mean of the rate in 2020 that accounts for the shock and the previously modeled rate for 2022.

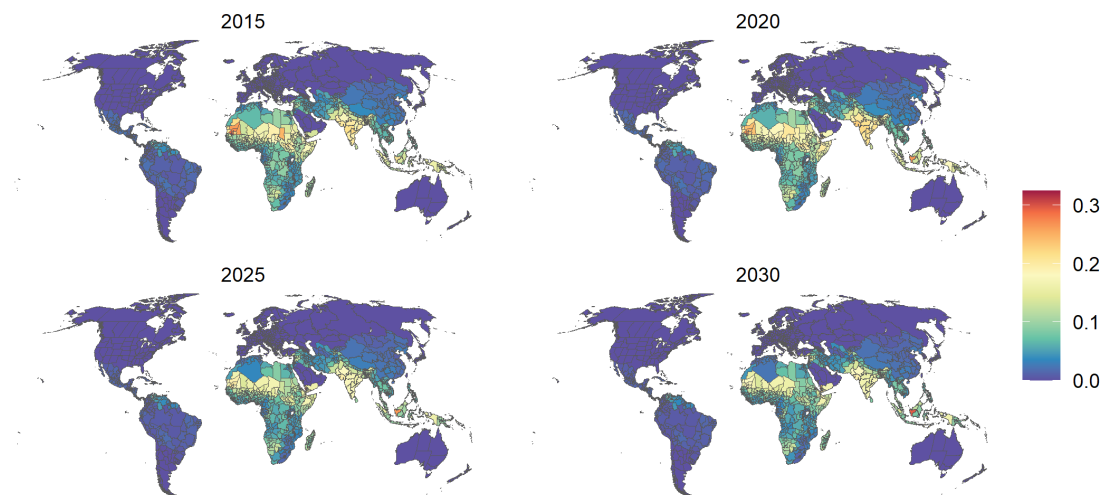


Figure S4: Prevalence of wasting.

S2.5 Stunting

We model stunting using the same methodology we use to summarize and model wasting, because the Local Burden of Disease dataset that we draw on for wasting also includes stunting. We also included the 14.3% increase in prevalence for the year 2020 that was estimated for wasting [6], even though stunting is a long-term consequence of malnutrition and will likely not be as readily observable in a population as wasting would be. However, we use stunting in our model not as a cause of food insecurity, but rather as a proxy for chronic conditions of hunger, which are exacerbated by the coronavirus pandemic. Thus, estimating an impact of the coronavirus shock on stunting is appropriate for our model.

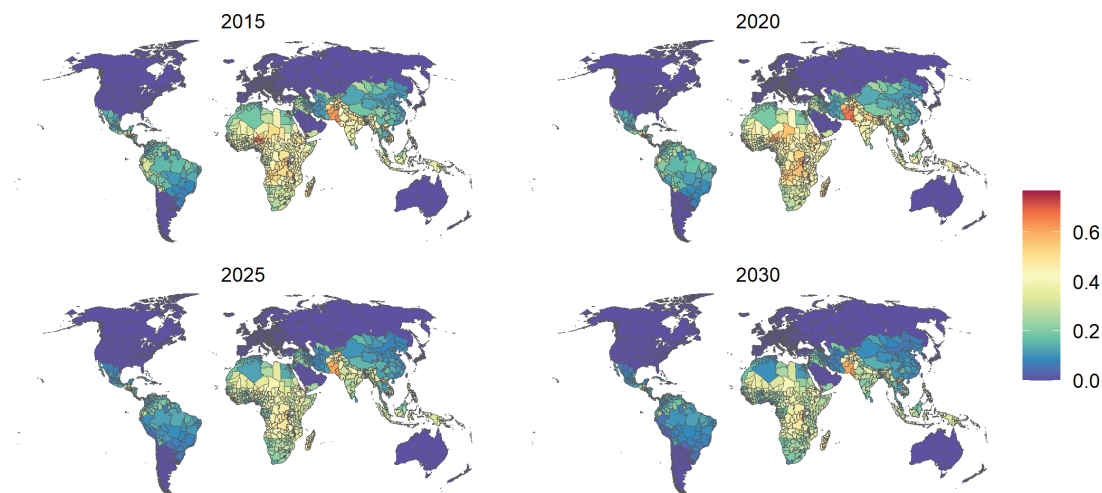


Figure S5: Prevalence of stunting.

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 patterns 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.

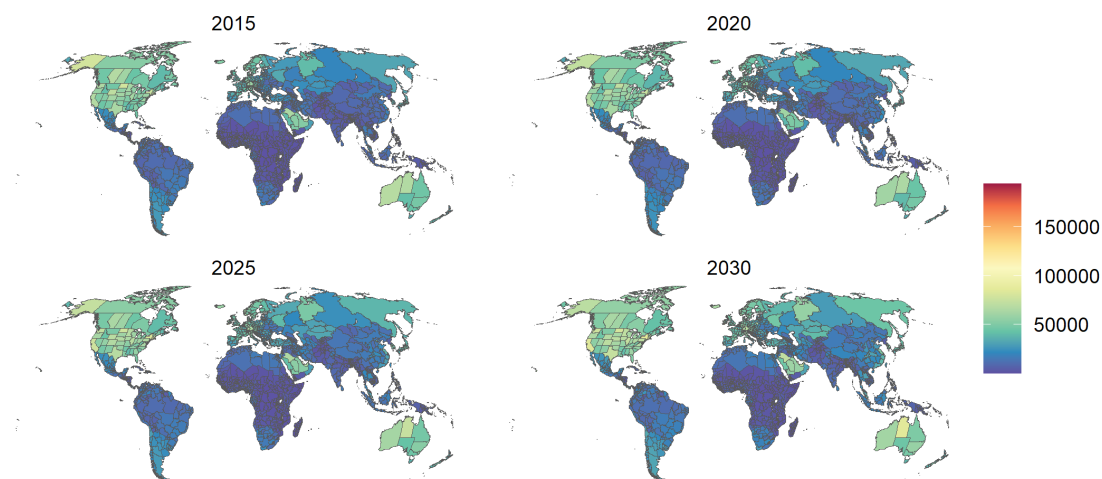


Figure S6: GDP per capita.

S2.7 Mean Years of Schooling

We estimate global subnational mean years of schooling using historic subnational estimates from the Subnational Human Development Database [2] and forecasts of national mean years of schooling from KC et al. [3]. Using the historic subnational data, we estimate each subnational administrative area's difference in years of schooling from the national mean, and use this value to disaggregate the future projections to a subnational level.

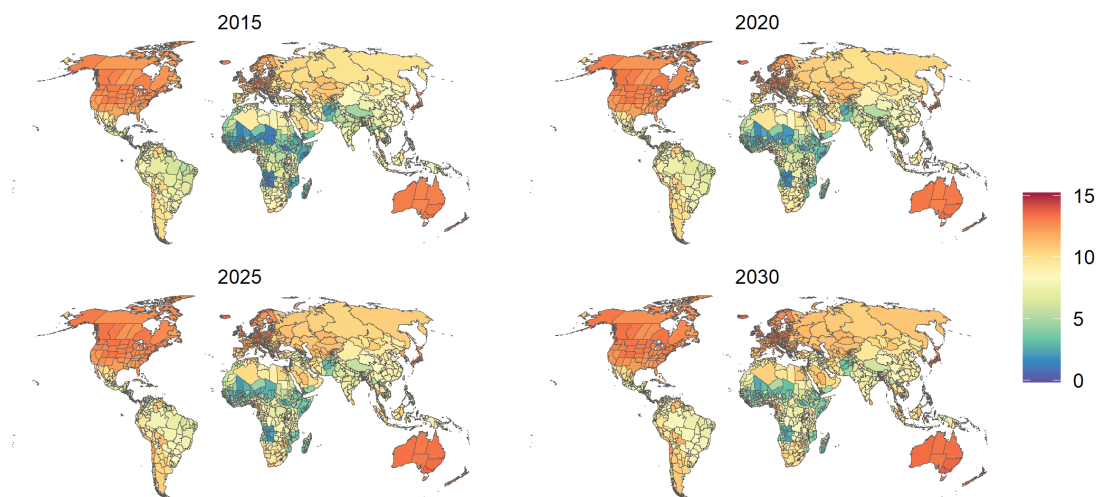


Figure S7: Mean years of schooling.

S2.8 Gini Coefficient

We use estimates of national income inequality from Rao et al. [9]. These estimates include all years in our analysis, and are not disaggregated to a subnational level.

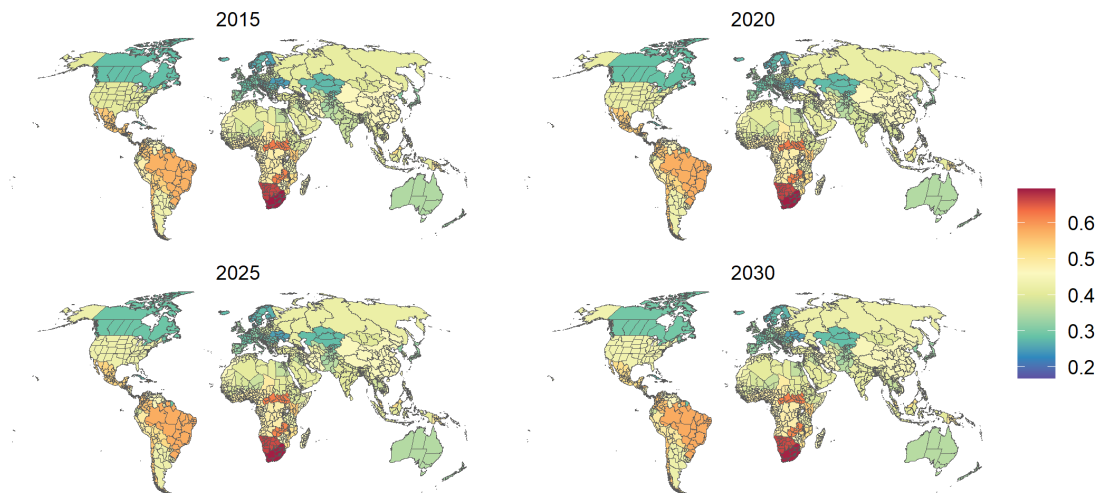


Figure S8: Gini Coefficient

S2.9 Poverty Headcount Index

We used estimates of the Poverty Headcount Index for the recent past and future from the World Poverty Clock by the World Data Lab [10], updated to include the impact of the coronavirus. In our model we include national level poverty rates at a threshold of \$1.90 which is defined as *extreme* poverty. By using the concept of parameterized Lorenz curves in combination with mean income/consumption and population projections future poverty rates are calculated. For more details on the methodology and information on global and country level poverty check the website worldpoverty.io.

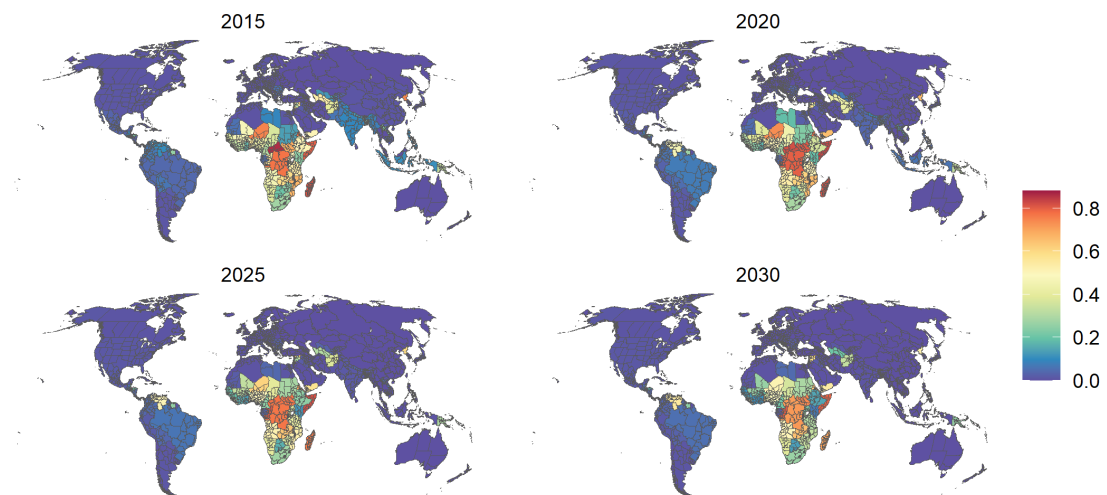


Figure S9: Poverty Headcount Index

S2.10 Water Scarcity Share

We used the water scarcity index (WSI) developed by [11]. The index represents the ratio of the decadal averages in water demand to water supply and it assesses changes in water scarcity on a 0.5 by 0.5 degree global grid. Based on this index, we estimate the share of people living in areas with less than 1000m³ of water available per capita per year. The WSI has been used already to build the water scarcity clock. For more information visit worldwater.io.

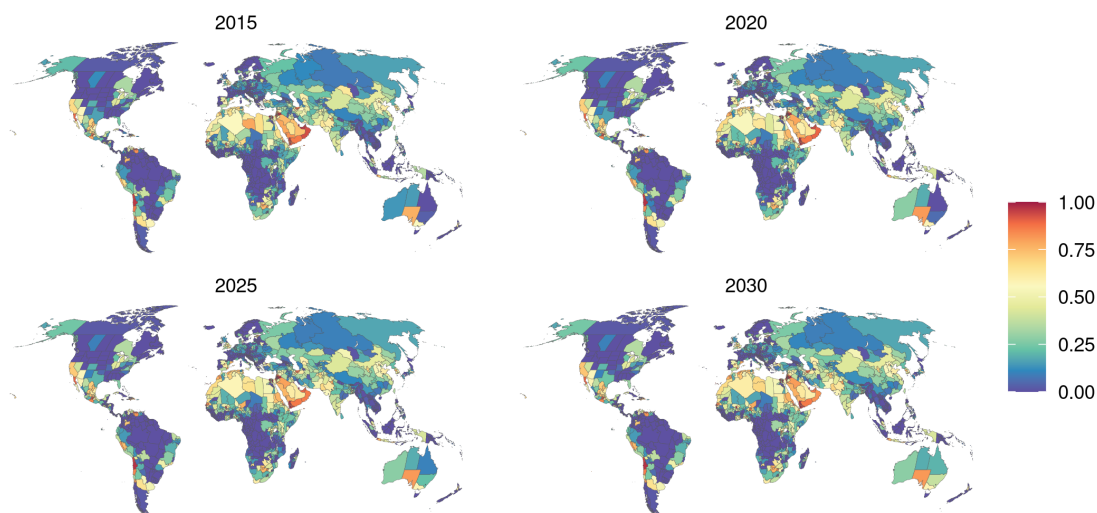


Figure S10: Water Scarcity Share

S2.11 Mean Annual Precipitation

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

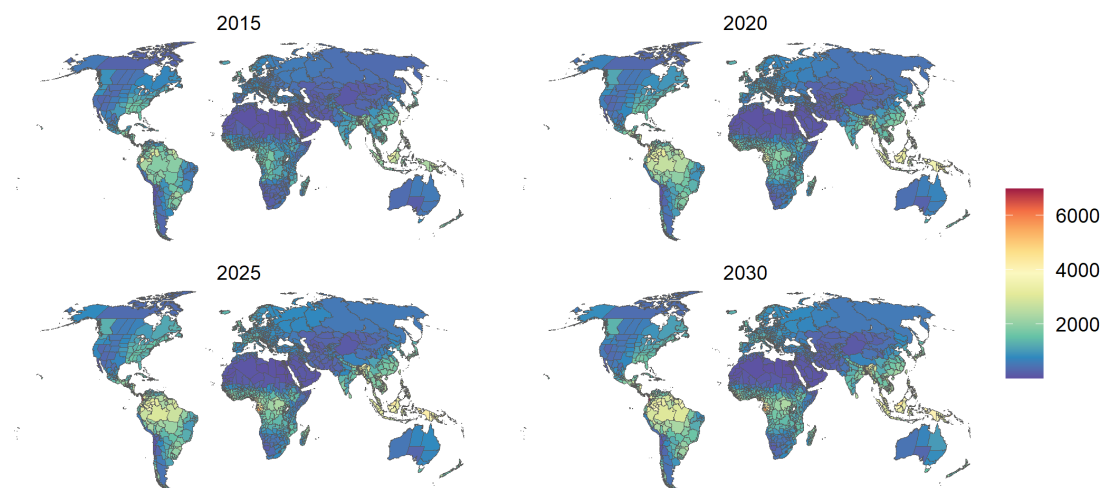


Figure S11: Mean annual precipitation

S2.12 Mean Temperature

For mean temperature, we used the same datasets that we'd used for precipitation, as each of those datasets included both temperature in addition to precipitation.

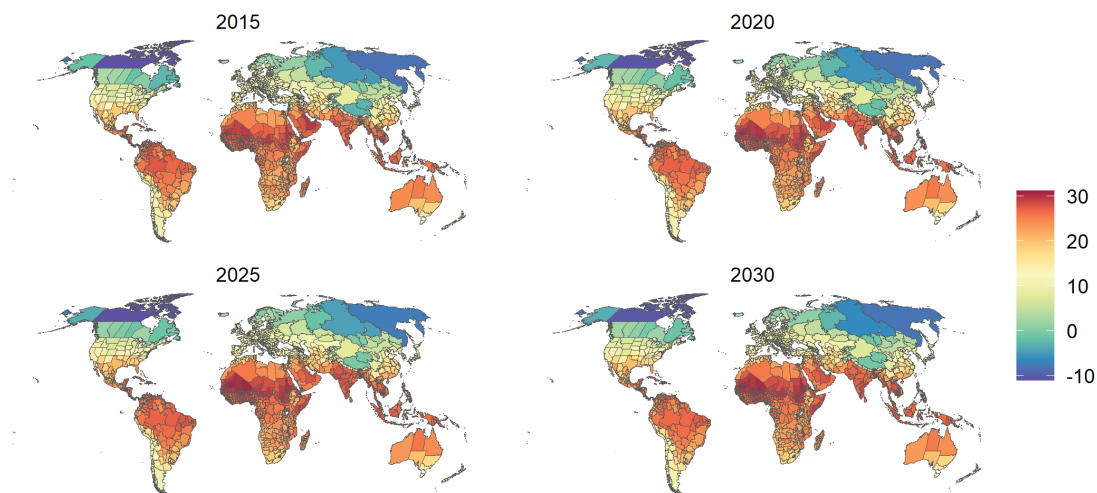


Figure S12: Mean temperature (Celsius)

S2.13 Topographic Ruggedness

As a measure of accessibility, we include the mean topographic ruggedness of each subnational area. We first used a gridded dataset of elevation from the USGS [14] and calculated the index at each grid cell using the methodology from Riley et al. [15]. We then summarized the values of all the grid cells within each administrative area. Because this variable does not change over time, it is constant across all years in the model.

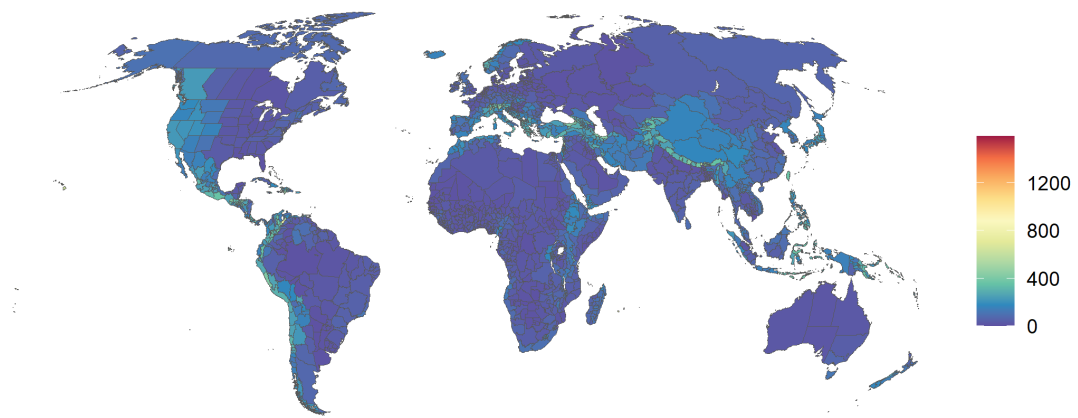


Figure S13: Topographic ruggedness

S2.14 Malaria (*P. falciparum*) Mortality Rate

We used data from *The Lancet* to estimate the mortality (deaths per 100,000 population per year) due to Malaria [16], taking the average pixel values within each subnational administrative area. We then estimated future malaria mortality based on AROC method.

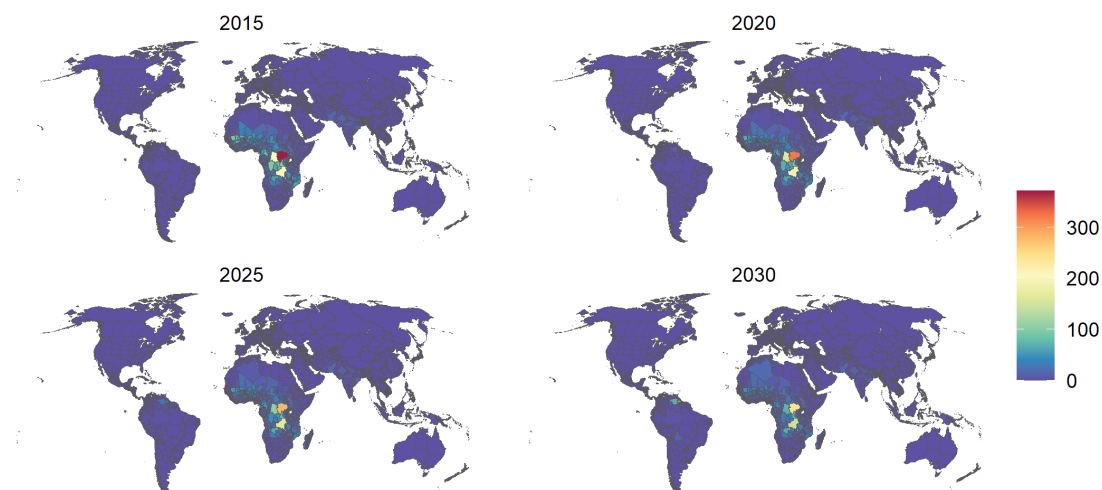


Figure S14: Rate of mortality due to *P. falciparum* Malaria

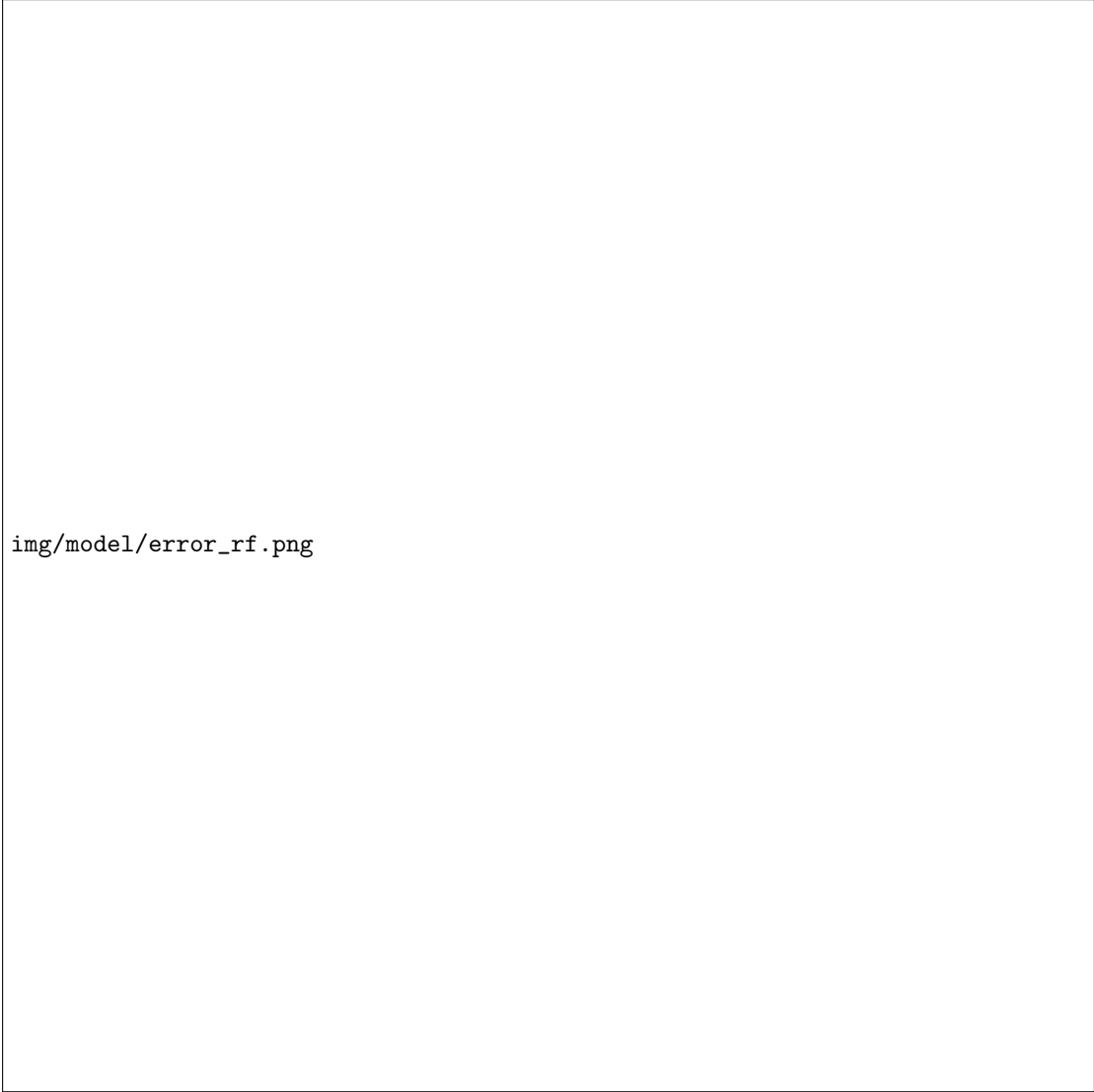
S3 Description, Implementation and Validation of the RF regression model

The RF regression is a machine learning method based on the creation of a large number of decision trees. By taking an ensemble of decision tree models, random forests introduce more variance and balance out the bias that is common to single decision trees [17]. 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 area also selected at random.

For implementation we use the R-package `randomForestSRC` by Ishwaran and Kogalur [18]. Specifically, we apply the functions `tune.rfsrc()` to find the optimal combination of hyperparamters and the function `rfsrc()` to perform the RF regression models. After predicting rates of moderate-to-severe and severe food insecurity we plot the modeled vs. observed rates and calculated the MSE and R^2 (See Fig. S16 and Fig. S17) for both models. Additionally, we use the function `vimp()` to gain insights into the development of the error rates with increasing numbers of trees (See Fig. S15) and the importance of individual variables (See Fig. S18).

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 ensure that predicted rates are bounded by 0 and 1 we conduct a logit transformation of the dependent variable and inverse logit transformation of the predictions.



img/model/error_rf.png

Figure S15: Error Rate of the RF regression model. Panel (A) shows the Moderate-to-Severe model and Panel (B) shows the Severe model.

In addition to the parameter that describes the number of trees to be created, two other important hyperparameters must be tuned. In the function `tune.rfsrc()` these parameters are called `mtry` and `nodesize`. `mtry` describes the number of variables randomly selected as candidates for splitting a node and `nodesize` describes the average number of observations in a leaf node. The optimal pair of hyperparameters is found by trying different combinations and choosing the one with the smallest out-of-bag (OOB) error.

The function `tune.rfsrc()` does exactly that and returns optimal values for `mtry` and `nodesize`. We applied this function on both, the moderate-to-severe and the severe model. For the moderate-to-severe model the optimal combination is `mtry = 12` and `nodesize = 1` and for the severe model we found `mtry = 10` and `nodesize = 2` to be optimal.

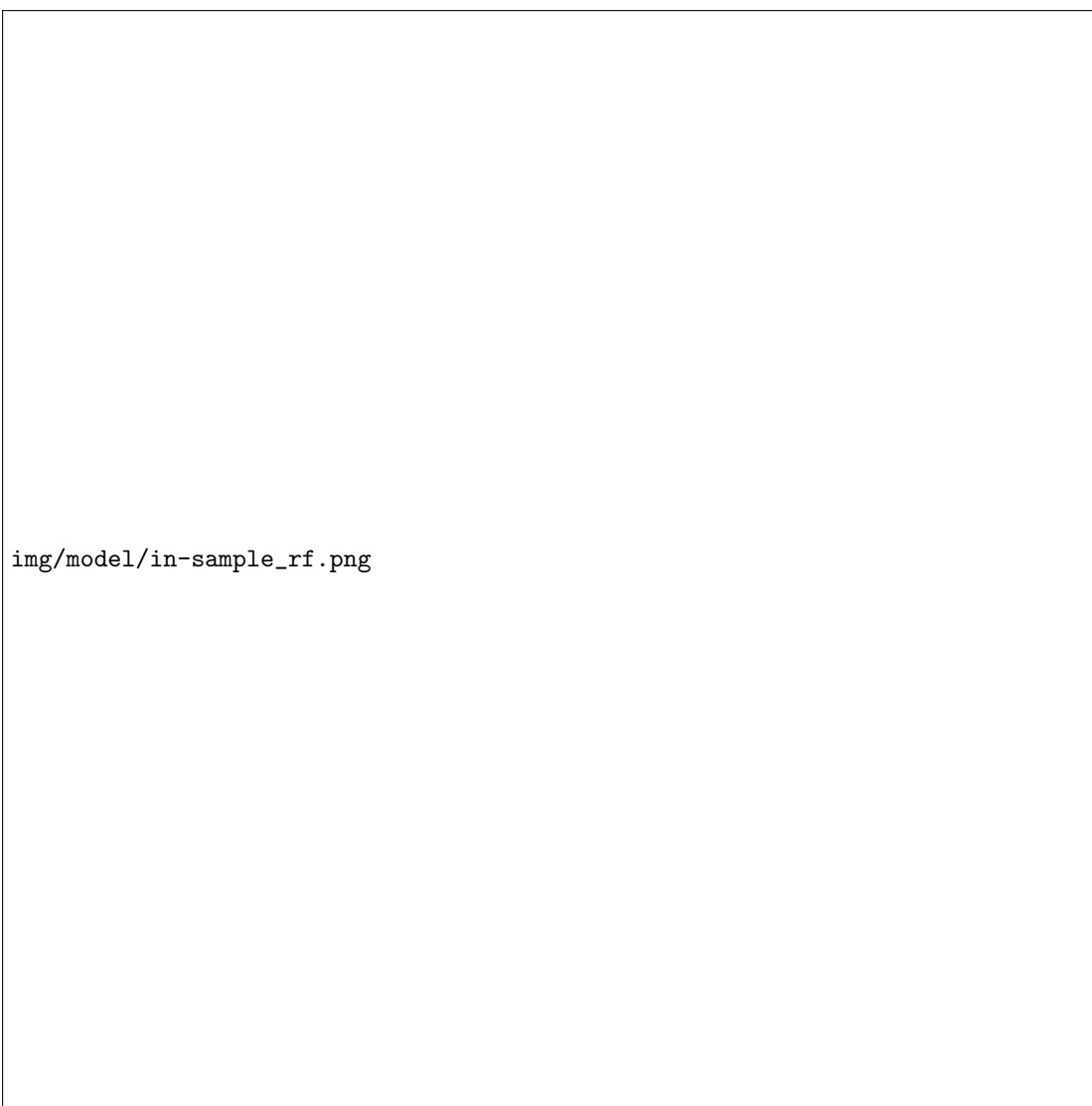


Figure S16: In-Sample Fit of the RF regression model. Panel (A) shows the Moderate-to-Severe model and Panel (B) shows the Severe model.

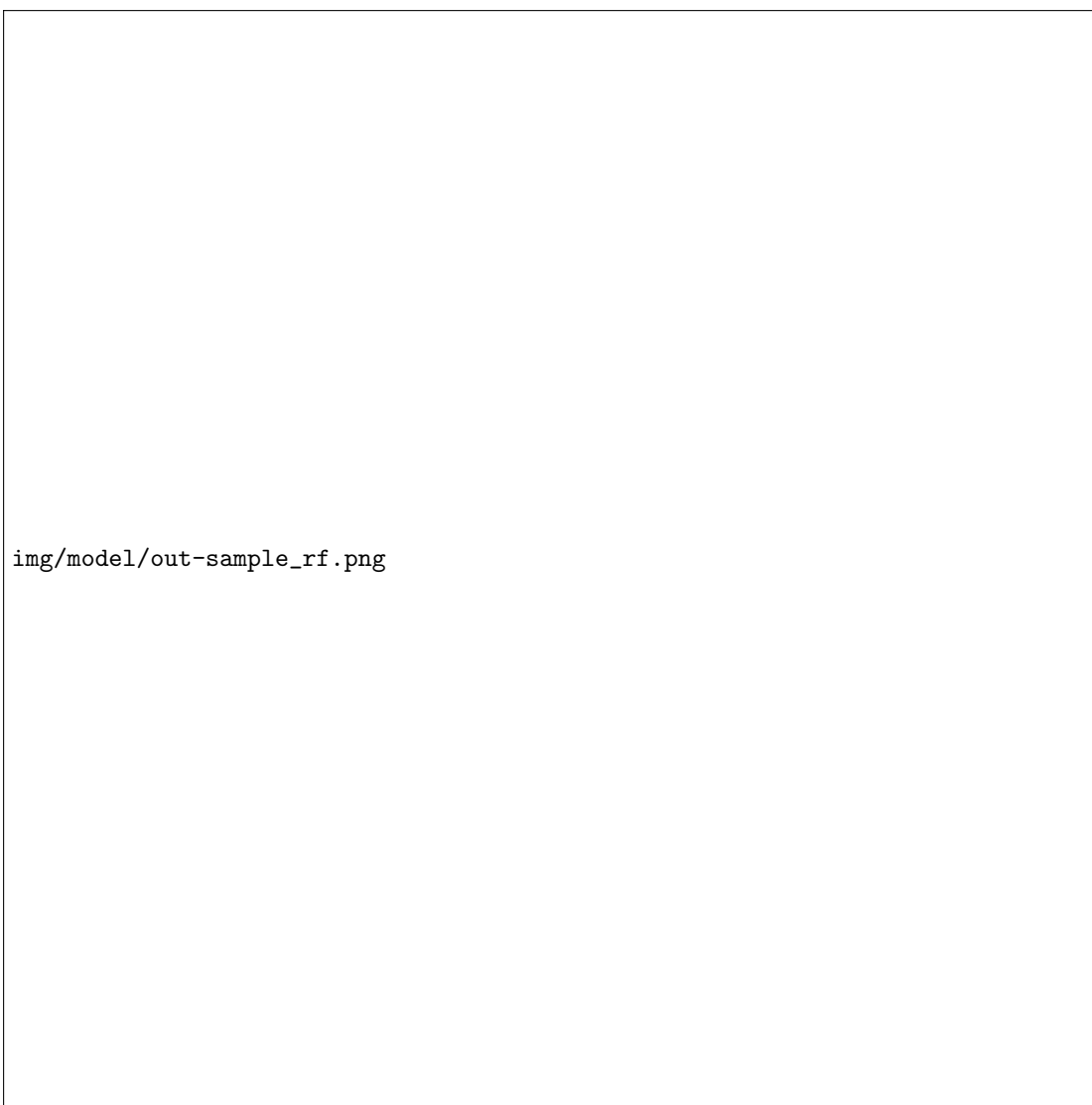


Figure S17: Out-Of-Sample Fit of the RF regression model with 75% training set and 25% test set. Panel (A) shows the Moderate-to-Severe model and Panel (B) shows the Severe model.

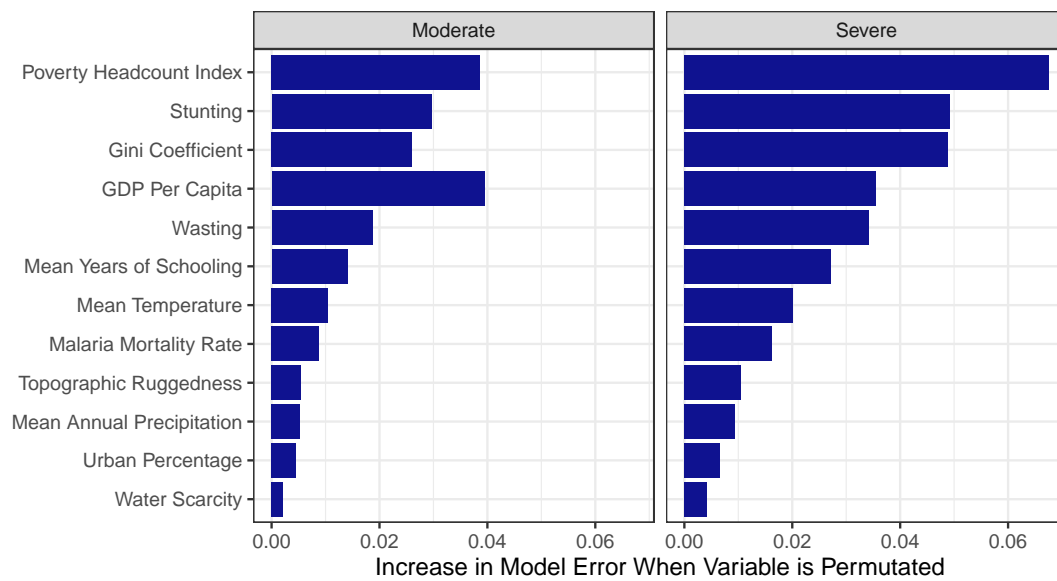


Figure S18: Variable Importance of the RF regression model. Panel (A) shows the Moderate-to-Severe model and Panel (B) shows the Severe model.

Lastly we assess the importance of individual variables. This is based on the Breiman-Culter permutation variable importance information [19]. This involved comparing the prediction error on the OOB data to the prediction error of OOB cases where the x -variable x is randomly permuted. The importance of an individual variable is thus the mean difference between the perturbed and unperturbed error rate, across all trees.