NBER WORKING PAPER SERIES

GLOBAL HIGH-RESOLUTION ESTIMATES OF THE UNITED NATIONS HUMAN DEVELOPMENT INDEX USING SATELLITE IMAGERY AND MACHINE-LEARNING

Luke Sherman
Jonathan Proctor
Hannah Druckenmiller
Heriberto Tapia
Solomon M. Hsiang

Working Paper 31044
http://www.nber.org/papers/w31044

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2023, Revised August 2025

This work was supported by a grant from the Human Development Report Office of the United Nations Development Programme. We thank Pedro Conceicao and seminar participants at The Workshop in Environmental Economics and Data Science, the WIDER Development Conference (Bogota), IDinsight, and the American Geophysical Union Fall Meeting for their valuable feedback. We thank the Global Data Lab for sharing DHS cluster-level data on the International Wealth Index. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Global High-Resolution Estimates of the United Nations Human Development Index Using Satellite Imagery and Machine-learning

Luke Sherman, Jonathan Proctor, Hannah Druckenmiller, Heriberto Tapia, and Solomon M. Hsiang

NBER Working Paper No. 31044
March 2023, Revised August 2025
JEL No. C1, C8, I32, R1

ABSTRACT

The United Nations Human Development Index (HDI) is arguably the most widely used alternative to gross domestic product for measuring national development. This is in large part due to its multidimensional nature, as it incorporates not only income, but also education and health. However, the country-level resolution of the global HDI data released by the Human Development Report Office of the United Nations Development Programme (N=191 countries) has limited its use at the local level. Recent efforts used survey data to produce HDI estimates for first-level administrative units (e.g., states/provinces). Here, we build on recent advances in machine learning and satellite imagery to develop the first global estimates of HDI for second-level administrative units (e.g., municipalities/counties, N = 61,530) and for a global 0.1° × 0.1° grid (N=819,309). To accomplish this we develop and validate a generalizable downscaling technique based on satellite imagery that allows for training and prediction with observations of arbitrary shape and size. This enables us to train a model using provincial administrative data and generate HDI estimates at the municipality and grid levels. Our results indicate that more than half of the global population was previously assigned to the incorrect HDI quintile within each country, due to aggregation bias resulting from lower resolution estimates. We also illustrate how these data can improve decision-making. We make these high-resolution HDI estimates publicly available in the hope that they increase understanding of human wellbeing globally and improve the effectiveness of policies supporting sustainable development. We also make available the satellite features necessary to increase the spatial resolution of any other administrative data that is detectable via imagery.

Luke Sherman
University of California, Berkeley
lsherman@berkeley.edu

Jonathan Proctor
University of British Columbia
Food and Resource Economics
jon.proctor@ubc.ca

Hannah Druckenmiller
California Institute of Technology
and NBER
hdruck@caltech.edu

Heriberto Tapia
Human Development Report Office,
United Nations Development Programme
heriberto.tapia@undp.org

Solomon M. Hsiang
Stanford University
and NBER
solhsiang@stanford.edu

Global HDI estimates and replication data/code are available at www.mosaiks.org/hdi

Global high-resolution estimates of the United Nations Human Development Index using satellite imagery and machine learning

Luke Sherman[1*], Jonathan Proctor[2*†], Hannah Druckenmiller[3], Heriberto Tapia[4], Solomon Hsiang[1]

[1]Global Policy Lab, Stanford University

[2]University of British Columbia

[3]California Institute of Technology

[4]Human Development Report Office, United Nations Development Programme

[*] equal contribution

[†] correspondence: jon.proctor@ubc.ca

August 2025

Abstract

The United Nations Human Development Index (HDI) is arguably the most widely used alternative to gross domestic product for measuring national development. This is in large part due to its multidimensional nature, as it incorporates not only income, but also education and health. However, the country-level resolution of the global HDI data released by the Human Development Report Office of the United Nations Development Programme (N=191 countries) has limited its use at the local level. Recent efforts used survey data to produce HDI estimates for first-level administrative units (e.g., states/provinces). Here, we build on recent advances in machine learning and satellite imagery to develop the first global estimates of HDI for second-level administrative units (e.g., municipalities/counties, N = 61,530) and for a global $0.1^{\circ} \times 0.1^{\circ}$ grid (N=819,309). To accomplish this we develop and validate a generalizable downscaling technique based on satellite imagery that allows for training and prediction with observations of arbitrary shape and size. This enables us to train a model using provincial administrative data and generate HDI estimates at the municipality and grid levels. Our results indicate that more than half of the global population was previously assigned to the incorrect HDI quintile within each country, due to aggregation bias resulting from lower resolution estimates. We also illustrate how these data can improve decision-making. We make these high-resolution HDI estimates publicly available in the hope that they increase understanding of human wellbeing globally and improve the effectiveness of policies supporting sustainable development. We also make available the satellite features necessary to increase the spatial resolution of any other administrative data that is detectable via imagery.

Introduction

The Human Development Index (HDI) is widely used by policymakers and academics to summarize three key dimensions of wellbeing: the population's health, human capital, and standard of living (*1–4*). A more comprehensive measure of wellbeing than income or wealth alone (*2*, *3*, *5*), HDI is used to categorize countries by their level of human development, which, in turn, can determine allocations of global resources (*6*). However, the United Nations Development Programme (UNDP) releases official global estimates of HDI annually only at the highly aggregated national level (N=191), preventing the use of the indicator in applications that require subnational information. Thus, measures of income remain the dominant metric for evaluating development progress within countries, in part because they are more readily available.

In an effort to address this, non-UN researchers (*7*) recently processed extensive household survey data in order to produce the first HDI estimates for first-level administrative units – i.e. provinces and states, hereafter called "provinces" (N=1,739). These efforts have substantially advanced our understanding of global development patterns, but province-level measures nonetheless remain too coarse for many modern policy applications where local information is needed, such as community-level targeting of aid (*8*, *9*). Furthermore, the reliance of current HDI estimates on slow, infrequent, and costly ground-based data collection limits the usability of HDI for most practical applications other than cross-national rankings.

Here, we produce the first global estimates of HDI at the second administrative level – i.e. municipalities and counties, hereafter called "municipalities" (N=61,530) and for a global $0.1^\circ \times 0.1^\circ$ (approximately 10km by 10km) grid. We construct these estimates by combining information from prior provincial estimates (*2*) with global daytime and nighttime satellite imagery (*10*, *11*). Our approach builds on recent advances in machine-learning (*12*) to develop a general method that learns the relationship between imagery and an outcome of interest (here, HDI) using data from any set of political boundaries. We can then use that relationship to estimate the outcome for any other set of boundaries. Importantly, our method works for spatial units of arbitrary shape and size, so models can be trained on coarse-resolution outcome measurements and make predictions at finer resolution. We apply this method to transform provincial HDI measures into finer resolution estimates. While other such "downscaling" approaches typically rely on either theoretically informed relationships or simple dasymetric masks (*13*, *14*) our approach to making predictions at finer resolution than the source labels uses machine learning and satellite imagery to identify complex spatial relationships between imagery and an outcome of interest.

A general approach to downscaling administrative data using satellite imagery and machine learning

The combination of satellite imagery and machine learning (SIML) is increasingly used to predict socioeconomic variables at fine spatial resolution (*12*, *15–22*). This approach enables information that is expensive to obtain through ground surveys to be estimated at low cost. While SIML estimates do not replicate ground surveys exactly (*23–25*), the quality of SIML estimates is now high enough that it can assist targeting of aid and program evaluation in remote communities where alternative sources of information are unavailable (*8*, *15*, *24*, *26*, *27*).

However, the ability of SIML systems to promote development is limited by the paucity of suitable observations for model training (*15*). This limitation is partly due to the design of modern SIML methods, since large quantities of administrative data are available, but existing systems are generally not designed to make use of them. To date, SIML approaches for predicting human outcomes have standardized the structure of both the training labels and corresponding imagery so that the unit of analysis is a regular spatial structure, such as a square. For example, many systems use convolutional neural networks (CNNs)(*16*, *17*, *20*), which tend to perform well on diverse computer vision tasks. CNNs, however, typically require images to be a constant size and shape, such as 224 x 224 x 3 pixels in the case of the commonly used ResNet-18 (*28*). This restriction has caused prior studies to rely on coarse approximations for linking irregularly shaped labels to corresponding imagery, for example, by averaging polygon labels that overlap with the square image (*12*, *18*). Such procedures can introduce considerable error when administrative polygons are much larger or smaller than the chosen square size. This is particularly relevant for HDI, for which data is globally available only for nations or provinces, which tend to be irregularly shaped and vary greatly in spatial extent. For example, the largest provincial polygon in our data is the Far Eastern Federal District of Russia, which is over 6 million km^2, and the smallest is Banjul of Gambia, which is 7 km^2. Developing a robust and widely applicable SIML system that can be trained on inputs that correspond with such diverse administrative structures requires an alternative strategy.

In an ideal setting, we would solve for a function that could directly map a single satellite image "tile" (e.g. 1km $\times$ 1km) to the corresponding HDI for the same tile

$$HDI_{tile} = f(satellite\_image_{tile}) + \epsilon_{tile} \tag{1}$$

where ϵ is the component of HDI that is not measurable with imagery. In theory, Eq. 1 could be solved directly with many learning approaches, such as a CNN (*29*), but this is

infeasible in practice because tile-level data on HDI (i.e. the left-hand side of Eq. 1, HDI_{tile}) does not exist. Instead, we observe only aggregated estimates of HDI over politically-defined regions ($HDI_{country}$ or $HDI_{province}$) that correspond with large and irregular agglomerations of image tiles. For the SIML system described by $f(\cdot)$, this creates a mismatch between the spatial structure of inputs (image tiles) and outputs (administrative regions).

We solve this problem by converting image tiles into a generalizable set of descriptive variables or "features," X_{tile}, such that $f(\cdot)$ can be structured as linear in these features,

$$f(satellite\_image_{tile}) = \beta \cdot X_{tile}, \tag{2}$$

where β is a vector of weights (i.e. coefficients). Specifically, we construct a basis for the imagery such that outcomes of interest are well-represented by linear combinations of the basis vectors. This allows aggregate administrative measures of HDI to project onto corresponding aggregations of tile-level features with the same weights that would be recovered if the problem had been solved using only tile-level data. Thus, we learn the model

$$HDI_{province} = \beta \cdot \underbrace{\left(\frac{1}{N} \sum_{tile \in province} X_{tile} \right)}_{\bar{X}_{province}} + \epsilon_{province} \tag{3}$$

and recover the same weights β that we would have recovered had we directly solved Eq. 1 using the linearization in Eq. 2. See Supplementary Information S5 for an empirical validation of the scale-invariance of model weights using this approach. Note that $\bar{X}_{province}$ is simply the vector of average tile-level features within a province. The weights β can then be used to generate predictions for arbitrary aggregations of tiles. We use these β to downscale HDI to the municipality level ($\beta \cdot \bar{X}_{municipality} = \hat{HDI}_{municipality}$) and the tile level ($\beta \cdot \bar{X}_{tile} = \hat{HDI}_{tile}$).

The benefits of linearizing this problem have been understood in general terms, since linear models of basic scalar image properties (e.g. "greenness"(*30*) or nighttime lights (*31*)) have been widely used to downscale administrative-level data. However, to our knowledge, it has not been shown that such linearization is possible and skillful for the types of featurizations that capture complex spatial structures in imagery.

Here we demonstrate that such a skillful linearization can be achieved by embedding rich image information using the Multi-task Observation using Satellite Imagery and Kitchen Sinks (MOSAIKS) approach (*12*). Converting images into MOSAIKS features ($daytime\_satellite\_image \rightarrow \mathbf{X}_{MOSAIKS}$) provides a structured representation of the unstructured information within the satellite image that performs well in linear models. These

features summarize the joint distribution of color and textures within daytime tri-band optical imagery (see Methods 1 and Figure S1). We concatenate these features with features that summarize nighttime lights of locations (*32*, *33*) (*nighttime_satellite_image* $\rightarrow \mathbf{X}_{NL}$) to construct a linear model that downscales HDI using only satellite data (see Methods 3.2 and 4.1).

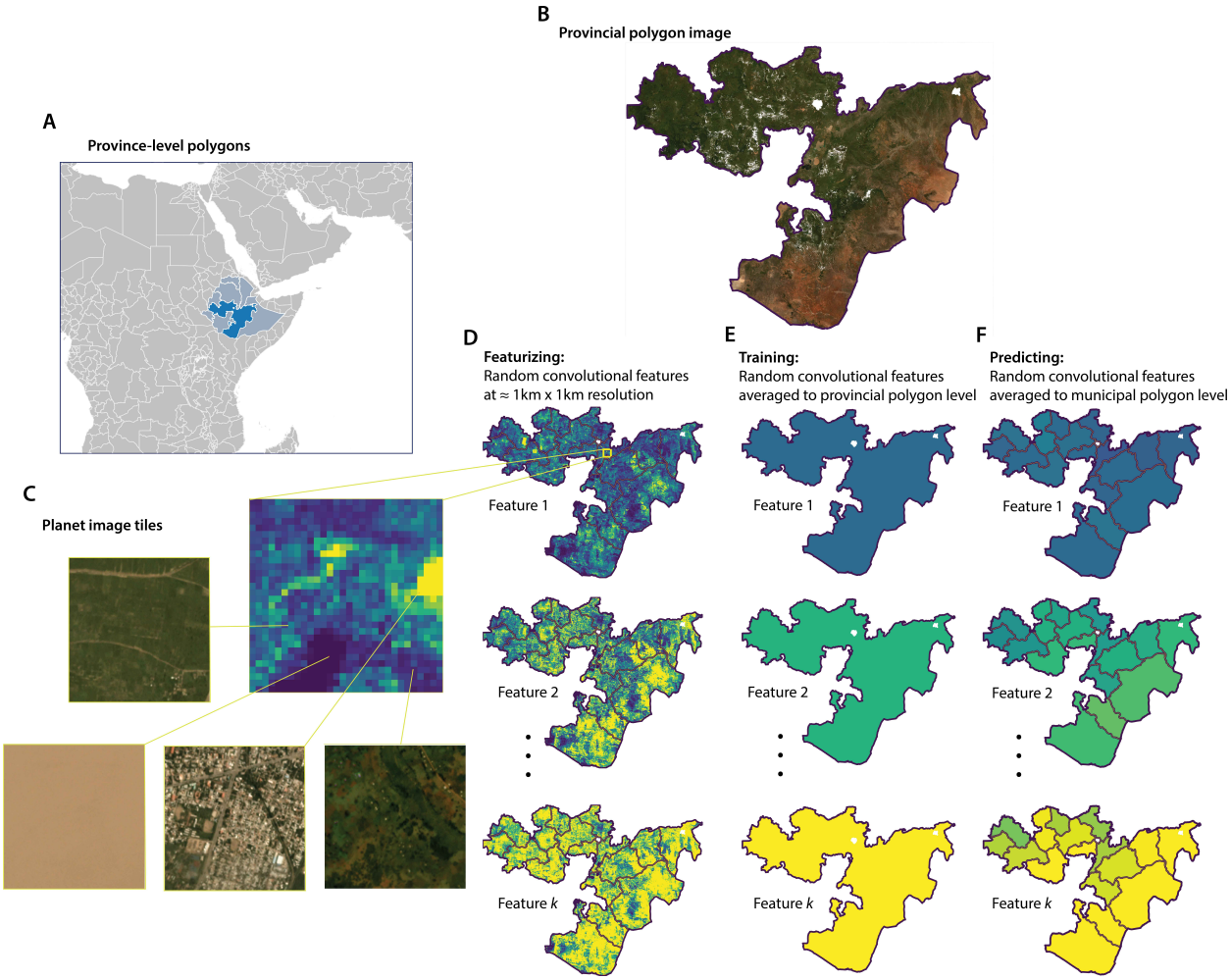

Figure 1: **The MOSAIKS approach transforms satellite imagery for each administrative polygon into a vector of image features**. (**A**) The location of Oromia, an example province (ADM1 unit) within Ethiopia. (**B**) A composite of Planet imagery over Oromia in 2019. (**C**) A sample of 0.01° x 0.01° image tiles. (**D**) Three examples of MOSAIKS random convolutional features over Oromia; each pixel shows the feature value for a single 0.01° × 0.01° image (X_{tile}). (**E**) The corresponding aggregation of these MOSAIKS features to the provincial polygon (ADM1) level for model training ($\bar{X}_{province}$). (**F**) Aggregation of these same MOSAIKS features to the municipal polygon (ADM2) level for fine-resolution prediction of HDI ($\bar{X}_{municipality}$). See Figure S1 for an illustration of how MOSAIKS features are calculated and used to predict HDI.

Results

Our results have four sections. First we train and evaluate a global model for HDI at the province-level using aggregates of satellite features. Second, we implement multiple tests to validate that this model is skillful. Third, we generate the first high-resolution global HDI data using this procedure, and we evaluate how these estimates compare to existing aggregated estimates. Fourth, we illustrate how these new high-resolution estimates could alter decision-making when targeting aid.

1. Predicting province-level HDI using satellite imagery

Using province-level administrative HDI data for training (as in Equation 3 and Figure 1), we find that predictions made using linear aggregates of daytime and nighttime satellite image features, anchored to known country means, explain 96% of the variation in global provincial HDI values (Figure 2A, denoted "full variation performance"). Specifically, we train a model to predict provincial HDI deviations from each country mean and then add the known country mean to the predicted provincial deviations. We take this "mean-anchored" approach, because it reflects how SIML may be used to augment existing HDI data in practice (see Methods 4.4 and Supplementary Information S6.1 for a discussion of mean-anchoring). Since most HDI variation is across-countries rather than within-countries (Figure 2A-B), much of this performance predicting provincial variation is driven by the measured country means that the predictions are anchored to.

The primary value of incorporating satellite imagery is to explain *local-scale* variation in HDI. As a first test of the model's ability to explain local variation, we evaluate model performance predicting deviations in provincial HDI from the country mean. We find that model predictions explain 52% of this *within-country* provincial variation in HDI (Figure 2B; see Methods 4.2 for a discussion of performance metrics). This indicates that SIML-based provincial predictions of HDI add substantial fine-resolution information to existing national measures. Importantly, models trained on provincial deviations from the country-level HDI have higher performance predicting such deviations than models trained directly on the provincial values themselves (Table S1 col. 4). Intuitively, model weights are optimized to explain the smaller within-country variation in the demeaned model, rather than the larger across-country deviations (Figure 2 A-B). We use such "within-country" models as our primary model specification.

2. Validating downscaling of data below the province-level

We cannot directly evaluate the performance of municipality-level or grid-level HDI predictions worldwide because such highly-resolved estimates have not been previously constructed. Nonetheless, we test the performance of our downscaling technique in three ways that allow predictions to be directly compared to "ground truth" at finer resolution than the training data. First, we directly compare our municipality-level HDI predictions to census-derived estimates in three countries where these data are available – Indonesia, Brazil, and Mexico. (*9*, *34*, *35*). Second, we train a model relating satellite imagery to the International Wealth Index (IWI) at the province level, and then construct downscaled predictions of IWI at the resolution of Demographic and Health Surveys (DHS) clusters where granular IWI measurements are available (*36*). The IWI is an alternative development indicator to HDI that omits measures of education and health. Third, we train a model to predict nighttime lights (NL), a common proxy for economic wellbeing (*32*, *37–42*), using features constructed exclusively from daytime satellite imagery, and test whether our approach can downscale NL. Mirroring the structure of our HDI analysis, we train a model using only NL labels aggregated to the province level, and then evaluate predictions of NL at the municipality level. No test can directly validate the performance for downscaling HDI globally, since the data necessary for such a test do not exist; however, all three of these large-scale tests taken together document the effectiveness of our downscaling strategy in general – using global socioeconomic data similar to HDI – and for HDI in particular – using municipal HDI data for three countries.

When evaluating estimates made at finer resolution than that of the training data, we mean-anchor downscaled estimates to the known provincial mean. This approach produces the best possible estimates by using the satellite-based model to explain within-province variation, which is previously unknown, and the known provincial values to explain the across-provincial variation.

Downscaling HDI in Mexico, Brazil and Indonesia As a direct evaluation of HDI downscaling performance, we compare municipal HDI predictions from the satellite-based model trained on provincial HDI deviations from the country mean to municipal HDI derived from census-based calculations in Mexico (*9*), Brazil (*35*), and Indonesia (*34*) (Methods 2.3). In Mexico, downscaled HDI predictions explain 45% of the municipal HDI variation overall (Figure 2C) and 29% of the within-province variation (Figure 2D). In Brazil, HDI predictions explain 48% of municipal HDI variation overall, and 20% of the within-province variation. And in Indonesia, HDI predictions explain 61% of the municipal HDI variation overall and 53% of the within-province variation. It is encouraging that HDI predictions align better with census-based measures for Indonesia than for Mexico or Brazil because measures

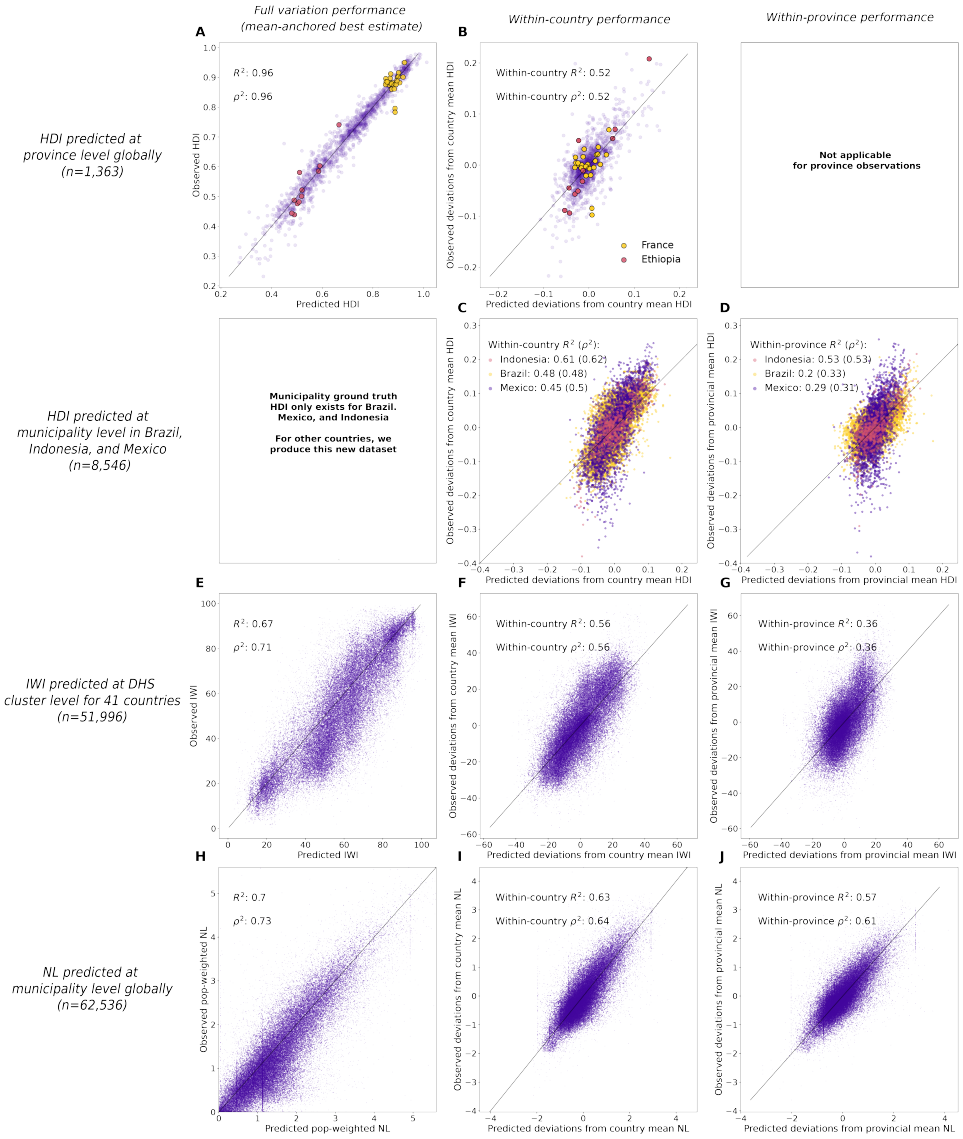

Figure 2: **MOSAIKS models perform well predicting socioeconomic indicators, including at downscaled resolution.** (**A**) Observed and predicted HDI at the province level. Note that the within-country variation is smaller than the across-country variation, as illustrated by France, in yellow, and Ethiopia, in pink. (**B**) The same as (**A**), evaluating within-country variation. Provincial deviations from the country mean for France and Ethiopia are now centered at 0, and the model is evaluated on how well it can differentiate provinces that are relatively well and worse-off within countries. (**C-D**) Observed and predicted municipal HDI data in Mexico, Brazil and Indonesia. (**E-G**) Observed and predicted IWI at the DHS cluster level. (**H-J**) Observed and predicted nighttime lights at the municipality level. The vertical streaking in (**H-J**) are caused by countries that have very spatially dense municipalities (Supplementary Information S6.3). Predictions are anchored to known country or provincial means (Methods 4.4). All predictions are for the year 2019.

for Indonesia are from 2019, which aligns with our satellite-based predictions of within-province HDI variation, while measures for Mexico and Brazil are from 2010. These older HDI measures could differ from our satellite-based predictions in part due to changes in HDI since the measurements were taken. Additionally, some portion of the misalignment between satellite-based predictions and survey estimates likely arises as a result of errors in the survey data itself, a widely recognized issue (*15*) that we are unable to assess. Together, these results indicate that our method for SIML-based downscaling improves our understanding of the spatial distribution of HDI in these three example countries; although, it is not a complete substitute for survey-based estimates when such data are available.

Downscaling the International Wealth Index We test the ability of our approach to downscale IWI internationally by training a model on province-level aggregates of IWI and then predicting IWI across DHS clusters. This is a more difficult task than predicting municipality-level values and an equally difficult task as predicting at the $0.1° \times 0.1°$ grid-level, since DHS clusters tend to be even finer resolution than municipalities and about the same size as the HDI grid (DHS cluster $\approx$ 180 km^2, municipality $\approx$ 2,000 km^2, grid tile $\approx$ 120 km^2). Models are trained on 862 provincial observations within 85 countries and evaluated at 51,996 DHS clusters (Table S1). Analogous to our approach with HDI, models are trained on province-level deviations from country-level means and predictions are re-centered to match the observed province-level mean, as these values are known (see Methods S6.2). Downscaled IWI predictions explain 67% of the variation in IWI across all DHS clusters (Figure 2E) and 56% of the variation in IWI across DHS clusters within countries (i.e. of cluster deviations from the country mean, Figure 2F). Importantly, this approach is also able to predict 36% of the variation in IWI within the provincial units that it was trained on (Figure 2G). This result demonstrates the ability of our downscaling approach to generate skillful global-scale predictions at resolutions higher than the training data, and also its ability generalize to measures other than HDI.

Downscaling nighttime lights To further evaluate our approach in a global test, we train a model on aggregate provincial NL and evaluate predictions at municipal resolution. NL are not a direct measure of human welfare; however, they are generally correlated with income and other development indicators (*38–40*, *43*, *44*). NL have even been used along with population to construct development indicators correlated with HDI (e.g. (*41*)). NL are particularly useful here because they allow us to design a validation test where true sub-national values are known worldwide. Mirroring our HDI process, we train the model using provincial deviations from the country mean, and then construct municipal values as the pre-

dicted municipal deviations from the country mean plus the known country mean (Methods S6.3). Downscaled NL predictions capture 70% of municipal variation in NL globally (Figure 2H), 63% of the municipal variation within countries (Figure 2I), and, most importantly, 57% of the variation across municipalities within provinces (Figure 2J). These results further reinforce the ability of our approach to downscale global province-level data and underscore its generalizability to other non-HDI outcomes. Unlike the two downscaling experiments above, this experiment relies entirely on features generated using only daytime imagery (Figure S2).

Comparisons to municipal data on HDI, IWI, and NL indicate that models trained on provincial data can explain 20-57% of within-province municipal variation. Collectively, these three experiments demonstrate that our approach effectively combines coarse socioeconomic measurements with satellite data to produce skillful estimates at spatial resolutions finer than the province-level training data.

3. Additional evaluation of model performance

Model performance for components of HDI One motivation for using MOSAIKS features to downscale HDI is their ability to predict a diversity of ground-based measures. This is particularly relevant for predicting HDI, since it is constructed from components that capture human health, education, and income. To consider which components of HDI are best captured by our estimates, we retrain models to predict each component of HDI separately. We find that MOSAIKS models explain 92% (5%) of the full (within-country) variation in provincial life expectancy, 93% (51%) of mean years of schooling, 90% (27%) of expected years of schooling, and 97% (56%) of gross national income per capita (GNIpc) (Table S3). While the components of HDI do tend to be correlated (Table S4), these results indicate that instead of just capturing income, predictions of HDI using satellite imagery maintain the ability to capture multiple dimensions of human wellbeing. These results also help explain what aspects of human development satellite features are able to capture, indicating their ability to explain local variation in education and income as well as their difficulty predicting aspects of health. Predictions of HDI made from combinations of its individually predicted components perform nearly identically to the direct predictions of HDI used throughout this analysis.

Further model evaluation including model, model performance across global regions, and the value of combining daytime and nighttime imagery are discussed in Supplementary Information S2. While it remains untestable how accurately the approach can downscale HDI globally, we find that the approach can predict substantial global variation in HDI and its

components, and that it can downscale both HDI in three countries and variables related to HDI globally with accuracy.

4. Global municipality-level and grid-level estimates of HDI

We use our model for within-country HDI (from Results Sections 1–2) to estimate HDI for 61,530 municipalities and 819,309 $0.1^\circ \times 0.1^\circ$ grid tiles (Figure 3), the finest resolutions at which HDI has been estimated globally (Methods 4.5). We make these municipal and grid-level estimates of HDI publicly available for download at `mosaiks.org/hdi`. Estimates are available annually from 2012-2021. We also similarly produce and make available estimates of the individual components of HDI.

Our high-resolution estimates enable a substantially more detailed understanding of human development compared with national and provincial measures (Figure 3 A,B vs. C,D). Both municipal and grid-level estimates reveal within-province heterogeneity of HDI that was previously un-resolved (Figures S3, S4). The gridded HDI estimates tend to be higher along major roadways, especially at the intersection of roadways (Figure S5). Boarders, such as between Turkey, Georgia, Armenia, Azerbaijan, Iraq and Iran, are less apparent in the fine resolution estimates, indicating a greater continuity in human development across space than in the provincial maps. The wealthier city centers and poorer suburbs of capital cities such as Moscow, Russia and Antananarivo, Madagascar are also visible in the municipal and grid estimates, but obscured in the provincial estimates. The contribution of environmental features to human development is illustrated in eastern Pakistan and northwestern India, where human development is higher in the plains bordering the Indus River and its tributaries, and lower in neighboring deserts. Similarly, within Sonora and Sinaloa in Mexico, coastal areas show higher human development than inland regions. This local heterogeneity in HDI indicates that uniform assignment of HDI to populations based on their country or province of residence is inaccurate because it groups together populations with very different levels of human development.

We use our estimates to quantify the degree of aggregation bias that occurs when using only province-level estimates. Aggregation bias occurs here because small units (i.e. grids or municipalities) are assigned the HDI of a larger unit (i.e. a province), which does not reflect local conditions. For example, a small urban region, where HDI is high, embedded in a province that also contains large, less developed rural areas, will be assigned a HDI level that is too low when provincial measures are used. We quantify how frequently such mis-assignment occurs within countries by assigning populations to a quintile of HDI within their national HDI distribution based on provincial, municipal, or grid-level estimates. We

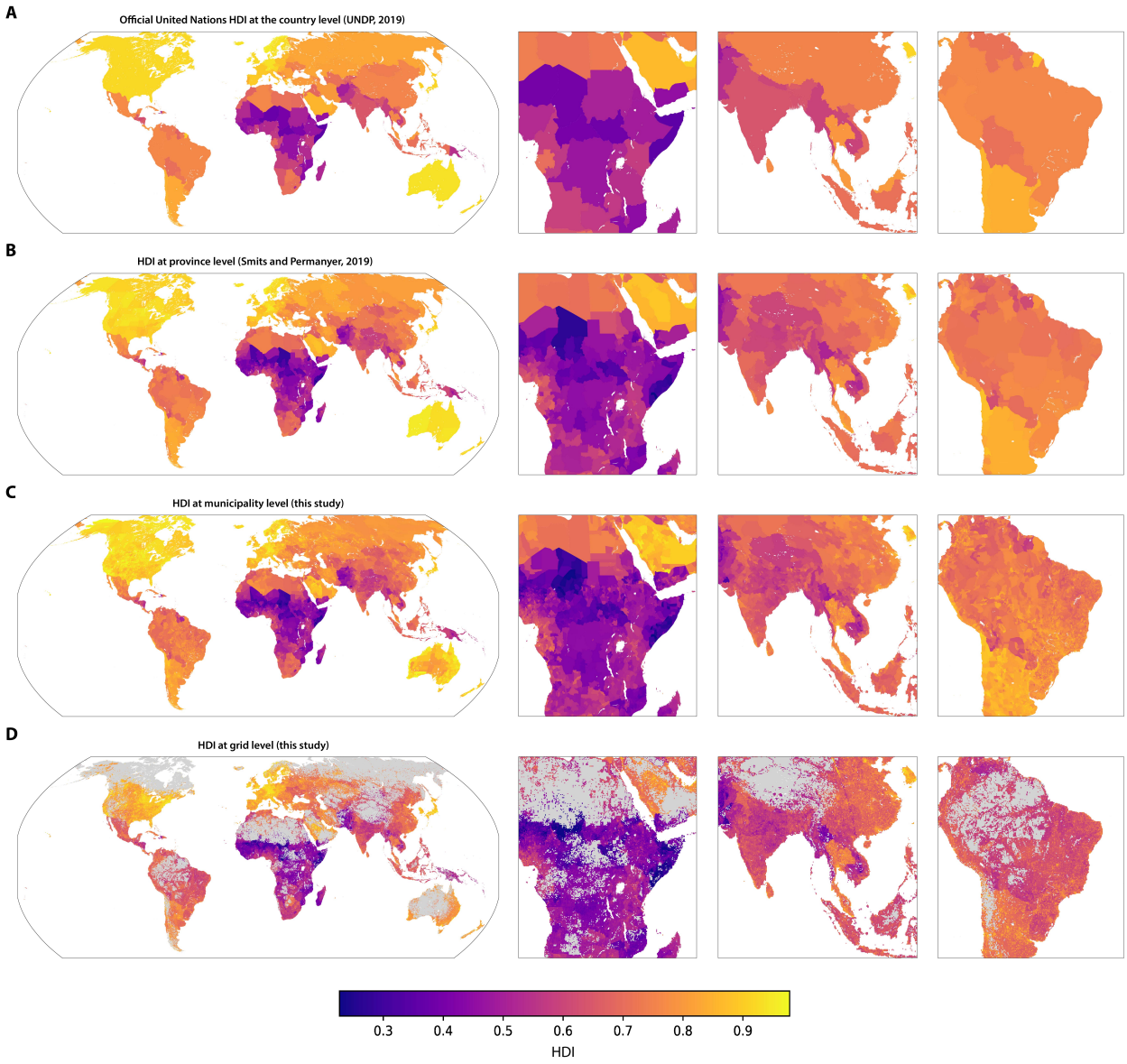

Figure 3: **Global HDI estimates at the municipal and grid levels.** (**A**) Official United Nations HDI at the country level (*45*) (**B**) HDI data at the province level from Smits and Permanyer (*7*). (**C**) Municipal level estimates of HDI produced here. (**D**) Grid level estimates of HDI at the 0.1° by 0.1° (approximately 10km by 10km) level produced here. Grey in the grid-level estimates indicates land area believed to be unsettled (*46*). All data shown are for the year 2019.

then evaluate how frequently the province-level estimates agree with the more highly resolved estimates (Figure 4).

We find that a majority of the global population (58% using municipal estimates and 65% using grid estimates) is assigned to a different within-country HDI quintile compared to when using provincial estimates. For example, of the population measured to be in the bottom two HDI quintiles using provincial estimates, 8.5% are reassigned to the top two HDI quintiles using municipal estimates and 12.9% using grid-estimates. Grid-level estimates reveal larger amounts of aggregation bias due to their finer resolution. Based on our grid-level estimates, we estimate that 20.4% (21.0%) of the global population is one quintile lower (higher) than assigned using provincial estimates, and 9.5% (9.3%) are two quintiles lower (higher).

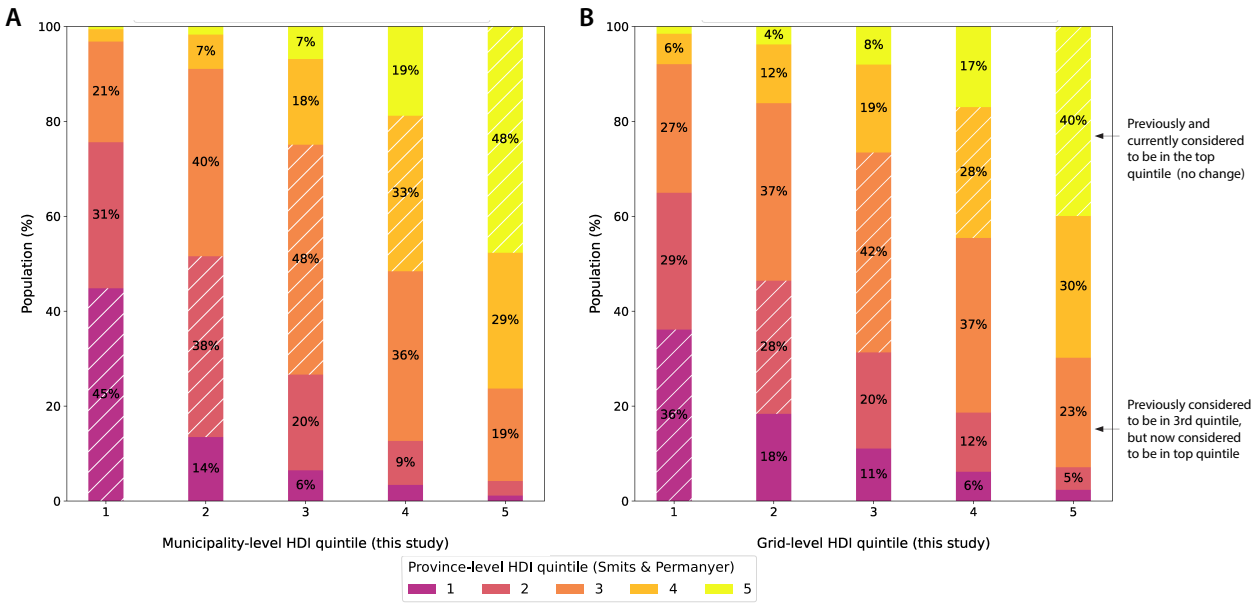

Figure 4: **Municipal and grid-level estimates of HDI assign more than half of the global population to a different within-country HDI quintile than provincial estimates.** (**A**) Shows the difference in estimated HDI quintile, within countries, using provincial vs. municipal data. (**B**) shows the same analysis using grid-level HDI estimates. Colors show the estimated HDI quintile using provincial data from (*7*), where yellow is high human development and purple is low human development. All data is from 2019. Bins along the x-axis show estimated HDI quintiles using the municipal and grid level data produced in this study. Hatch marks indicate no change in quintile assignment using municipal data. When provincial data do not allow for the creation of five distinct bins, the population is assigned first to the middle bin, followed by the neighboring bins. For example, if a country does not have any province-level data, the entire population is assumed to be in the middle quintile for that country. For this reason, a greater fraction of the global population is assigned to the middle quintile (33%) than to the outer quintiles when using the provincial data.

5. Illustrative application: targeting policy in Mexico

To explore how our new HDI measures could improve the efficiency of development policies, we conduct a simulation exercise for Mexico. We simulate how a geographically targeted policy based on provincial vs municipal HDI data (Figure 5A-B) might achieve different outcomes, noting that previous work has shown that more spatially granular targeting can produce meaningful welfare gains (*17*, *47–50*). We study Mexico because "ground truth" estimates of HDI enable us to benchmark performance (*9*).

The use of municipality-level data improves the number of program recipients correctly targeted. Supposing that the program director aims to provide assistance to the 10% of individuals with the lowest HDI, accuracy of program targeting increases by 11.4% percentage points (from 32.3% to 43.7%) when using municipal data (Figure 5E, assuming the standard deviation of individual HDI within each municipality is 0.1, see Methods 4.6). Use of the municipal data also results in a much greater geographic dispersion of targeted municipalities (Figure 5C-D). Evaluation of targeting performance using a receiver operating characteristic curve (*8*) gives similar results (Figure 5F, Methods 4.6). Replicating this analysis for Indonesia also gives consistent results (Figure S11).

This application illustrates how fine-resolution HDI estimates can improve targeting of policies that would otherwise use coarser provincial measures, even though these fine-resolution estimates are imperfect. Using provincial measures implicitly assumes no within-provincial variation in HDI, so any positive ability to predict variation within provinces improves understanding of the spatial distribution of HDI. Users of these fine-resolution estimates should consider their specific policy context to determine whether these fine-resolution estimates, which explain 29% of the within-province HDI variation in Mexico, 20% in Brazil, and 53% in Indonesia, provide sufficient additional information to be useful in their setting. These estimates may be particularly valuable in data poor settings, where traditional data sources such as surveys are less available and reliable (*15*); only about half of the world's poorest countries have conducted a census in the last decade (*16*, *26*). These estimates may also be useful in cross-country settings where a consistent metric of welfare is desired.

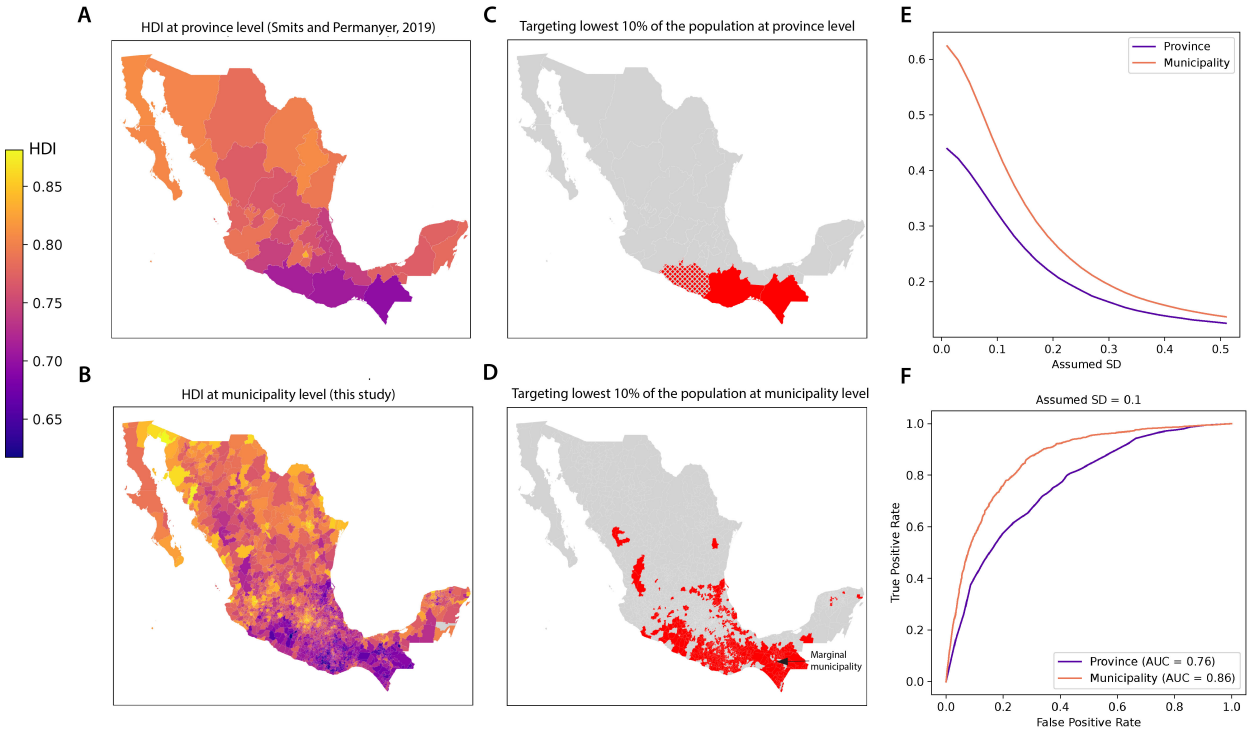

Figure 5: **Spatially granular HDI measures can improve decision-making.** (**A**) HDI for 2019 at the province level of observation (7) (**B**) HDI estimates for 2019 at the municipality level produced in this paper. (**C**) Lowest HDI provinces that would be targeted until 10% of the country's population is reached. (**D**) Lowest HDI municipalities that would be targeted until 10% of the country's population is reached. Hashing in (**C**) and (**D**) shows the marginal province and municipality that would be partially targeted. (**E**) Targeting accuracy (true positive rate) as a function of the assumed standard deviation of HDI within each municipality. (**F**) ROC curves illustrate the degree of improvement that comes with targeting at the municipality level relative to the province level (assumed SD of 0.1 within each municipality).

Conclusion

We produce and make publicly available the first global-scale, high-resolution estimates of HDI, enabling the use of broad-based measures of wellbeing for local decision-making and policy prioritization. We achieve this by developing an approach for generating spatially granular development indicators using SIML models trained on spatially aggregated and inconsistently structured administrative data. Since many forms of non-HDI data are also available only for administrative regions, we believe this generalizable method will increase the range of outcomes that can be used as labels in SIML models, which are currently constrained by limited training data (*15*).

Our strategy is motivated by the limited resolution of available training data for HDI, but our results do not exhibit obvious compromises in performance relative to alternatives that exploit high-resolution labels. A related benchmark in this literature achieves $\rho^2 = 0.63$-0.67 predicting a wealth index when training and evaluating at the DHS cluster resolution (*16*, *51*). Though we train at the provincial level, not the DHS cluster level, our performance predicting DHS cluster IWI is competitive with this previous analysis ($\rho^2 = 0.71$; Table S1, col 1). Another benchmark in this literature achieves an R^2 of 0.54 predicting RWI, which is conceptually identical to what we call within-country IWI, at the same DHS cluster level (*17*) (metric taken from a replication of their Figure 3B). Again, despite training at the provincial level rather than the DHS cluster level we achieve a within-country R^2 of 0.56 predicting IWI at the cluster level. Direct comparison to both of these benchmarks, however, is complicated by differences in training and evaluation methodologies including our focus on explaining within-province variation and our corresponding mean-anchoring approach.

Here, we use satellite imagery due to its rich information content, global availability, and near uniformity in quality across countries. To evaluate whether additional satellite data sources beyond visual daytime imagery and nighttime imagery might improve HDI estimates, we follow ref. (*22*) and add to the baseline model three additional index-based features constructed from Sentinel-2A satellite imagery that integrate knowledge of spectral properties of different ground conditions: the Normalized Difference Vegetation Index, Normalized Difference Water Index, and Normalized Difference Built-Up Index. We find that the model performance is essentially unchanged (Table S5), indicating that these additional features do not provide additional information beyond what is already captured by the MOSAIKS and NL features. While it is important to note that there is a limit to how well socioeconomic variables can be predicted using satellite imagery generally, and adding these additional features did not improve performance in this case, future work should nonetheless explore whether incorporating additional imagery sources and/or other ancillary data can

improve these estimates (*52*).

This approach to producing HDI estimates can be viewed as a type of regression-based technique for small area estimation using satellite imagery (*53*, *54*). Though data limitations in this setting – including the lack of fine-resolution global surveys of HDI – preclude the direct implementation of most commonly used small area estimation techniques (e.g., (*55*)), this approach follows the general small area estimation framework of combining limited measurements of the variable of interest (provincial HDI) with universally available axillary data informing the variable of interest (satellite imagery) to make accurate and precise fine-resolution predictions. A key difference between more traditional small area estimation techniques and ours is that we use a global-level auxiliary dataset that does not suffer from the challenge of construct validity across countries, although this comes with a tradeoff of needing to assume some degree of consistency between imagery and conditions on the ground. A recent comparison of poverty maps for Malawi produced using i) a small area estimation approach combining survey data and census data and ii) an alternative approach combining survey data and geospatial indicators including satellite imagery found that the two approaches produced very similar estimates, with a correlation exceeding 0.9 (*56*). This, paired with the performance of satellite imagery predicting HDI shown here, indicates that satellite imagery can be an effective substitute for census information when producing fine-resolution estimates of human wellbeing in settings where census data is not available. In locations where reliable census data is available, satellite imagery could serve as a complementary source of information.

One important property of the errors in our estimates, common in machine learning (*12*, *24*, *25*), is that our predictions exhibit lower variance than the true values. For example, in Mexico the standard deviation of our satellite-derived estimates is approximately half that of census-derived values. Survey and other traditional approaches to data collection remain critical to informing the state of global human development and complement satellite models, which cannot be trained or evaluated without ground-truth measures.

We have produced global downscaled estimates of HDI for 2012-2021 by combining existing provincial measures with fine-resolution satellite estimates. This approach estimates municipal HDI in each year by adding satellite-based estimates of municipal HDI deviations from the provincial mean, which are time-constant, onto time-varying provincial values calculated from surveys (*7*). These municipal values are appropriate for applications comparing levels of HDI across locations within a specific moment in time. They cannot be used to evaluate changes in the distribution of HDI within-provinces over time. Current data does not allow us to distinguish performance between our validated approach and an approach that estimates local trends in HDI (Figure S8 and Supplementary Information S3). Developing

additional approaches for tracking and validating subnational trends in HDI is an important area for future research.

A benefit of our approach is its mathematical transparency (*57*), relative to deep learning alternatives; however, like most computer vision approaches, it is a challenge to explain why our model generates a particular prediction. To explore which specific phenomena are correlated with our HDI estimates, we examine how predicted HDI values associate with other variables globally, an approach similar to some prior efforts to evaluate model predictions (*54*). We find that municipalities with higher road density (full variation ρ^2=0.24, within-country ρ^2=0.14) and building density (full variation ρ^2=0.07, within-country ρ^2=0.16) tend to have higher predicted HDI (Figure S9, Supplementary Information S4). Population has a weak association with HDI across the globe (full variation ρ^2=0.02), but within countries, more densely populated municipalities tend to have higher HDI (ρ^2=0.33). In contrast, cooler countries tend to have higher HDI (full variation ρ^2=0.12), but temperature has little association with HDI within countries (ρ^2=0.0). Terrain ruggedness, forest cover, crop cover, and rainfall have little to no association with our HDI estimates. Collectively this analysis indicates that our HDI estimates capture image information related to built infrastructure and the density of human settlements, but also indicates that there is residual image information beyond these factors that contribute to the HDI estimates – together the variables analyzed above explain 32% of the full variation in municipal HDI estimates and 33% of the within-country variation in a multiple linear regression analysis. Broadly, policymakers benefit from algorithmically-informed decisions that are transparent and explainable (*58*). Future work should investigate improvements in remote sensing model interpretability that can support transparent policy-making and aid in understanding potential sources of systematic bias (*57*, *59*, *60*).

We emphasize that the approach described here can be used to predict a wide variety of labels for which country, province, and/or municipality level labels exist. To facilitate this use, we make the features used in this analysis publicly available at the country, province, and municipality level via `https://mosaiks.org/hdi` (*10*). We offer these features aggregated to these administrative-unit levels using both area and population weights. Each of these files is relatively small, $\approx$ 3 GB or less, and thus possible to process on a desktop computer. For comparison, the global set of features is $\approx$ 3 TB and the raw imagery is $\approx$ 30 TB. We hope that researchers and decision-makers can leverage these features, along with their own administrative datasets, to produce new downscaled estimates of socially-relevant outcomes. Such spatially granular data may create new opportunities for achieving global development goals.

Methods

1 Overview

To transform satellite imagery into descriptive features that exhibit high performance in linear models, we build on the recent development of MOSAIKS, an approach that achieves performance competitive with CNNs using an unsupervised image embedding combined with a linear ridge regression model (*12*). The linearity of ridge regression enables the scale invariance of our downscaling approach (*12*). MOSAIKS random convolutional features combined with ridge regression have been shown to be skillful at solving diverse prediction problems —such as forest cover, population, elevation, and house price— using only imagery as inputs and using only a single linear specification. This property makes MOSAIKS a particularly appealing approach for predicting HDI, which is constructed from multiple development indicators. Each MOSAIKS feature for a tile describes the similarity between the satellite image and a smaller patch of imagery, and is calculated as a nonlinear transformation of the image's pooled convolution with a random sub-image from the sample (*12*, *61*). For example, a feature whose patch was sampled from a city might inform how urban an image is, while a feature whose patch was sampled from farmland might inform how agricultural an image is. Together, thousands of MOSAIKS features form a basis that can skillfully describe the rich structure contained within large imagery datasets through simple linear combinations of the features. For details on how MOSAIKS captures visual information from satellite imagery see Figure S1, as well as the Methods sections "Theoretical foundations" and "Convolutional random kitchen sinks", and the Supplementary Information section "Alternative interpretations relating MOSAIKS to kernels and CNNs" of ref. (*12*).

To compute local HDI via SIML, we transform a dataset of global Planet imagery ($\approx$ 5m resolution) into a set of 4000 general-purpose MOSAIKS features (X) for $0.01^\circ \times 0.01^\circ$ tiles ($\approx$ 1km $\times$ 1km; Figure 1) (*10*). We supplement these MOSAIKS features with features that flexibly characterize the distribution of nighttime lights in each tile (Methods 3.2). While past measures of economic and human wellbeing have tended to focus on nighttime lights alone (*37–42*), inclusion of high-resolution daytime imagery improves model predictions of HDI, especially in regions with low HDI (Supplementary Information S2). Visual imagery may improve performance by better resolving HDI variation in areas without electrification that are nearly dark, or by better differentiating between highly populated poorer areas and less populated richer areas, which can have similar brightness in nightlight data (*20*). We then learn a model that is linear in these features ($\beta \cdot X$) and use this linear model to estimate HDI at high resolution. While we focus this analysis on the downscaling of

HDI, this approach is generalizable to other types of administrative data associated with irregularly-shaped political units.

For both training and prediction, we average image features for administrative polygons using population weights (Figure 1D, grid-scale predictions follow a similar procedure) (*46*). This results in one vector of image features for each province and municipality in the world. To learn the relationship between the image features and HDI, we train a model on province-level HDI labels and aggregated province-level image features (Figure 1E). We then predict municipality-level HDI using the municipality-level image features (Figure 1F), and we predict $0.1^{\circ} \times 0.1^{\circ}$ grid HDI using features for that grid.

Throughout this analysis we use the term "province", the abbreviation "ADM1" to refer to first-level administrative regions; and "municipality", "ADM2" to refer to second-level administrative regions, though the terminology for these units varies by country. For example, "state" and "county" are the designations used for ADM1 and ADM2 units in the United States.

Our fine-resolution global HDI estimates were produced in collaboration with researchers at the Human Development Report Office of UNDP (Supplementary Information S1), but data released with this paper should not be considered official United Nations indicators.

2 Label data

2.1 HDI

The United Nations Human Development Index is a composite measure used to assess a country's average achievements in three key dimensions of human development: health, education, and standard of living (*4*). Health is measured by life expectancy at birth; education is evaluated using a combination of mean years of schooling for adults and expected years of schooling for children; and standard of living is assessed through gross national income per capita (GNIpc), adjusted for purchasing power parity. The HDI applies a logarithmic transformation to GNIpc to account for the diminishing benefits of increased income. Each of these components is normalized on a scale from 0 to 1, and the HDI is calculated as the geometric mean of the three dimension indices, giving equal weight to each dimension. This method allows the HDI to reflect not just economic wealth, but also broader aspects of well-being. In 2019, the average HDI across countries was 0.72, with a standard deviation of 0.15, a minimum of 0.36 in Somalia and a maximum of 0.96 in Switzerland. (Figure 3A).

2.2 Province-level HDI

National-level HDI data originate from the UNDP Human Development Data Center and are updated every year (*45*). UNDP uses data from the World Bank, UNESCO, UNICEF, DHS, UN Stats, and other organizations to create these national-level indicators (*4*).

Province-level data on HDI and its components come from the Global Data Lab (GDL) Subnational HDI Database V7.0 (*7*, *62*). We omit 3% of the observations, which do not match with the associated GDL shapefile. The resulting province-level HDI dataset contains 1,739 provincial observations from 159 countries. Additionally, we include 20 country-level observations that do not have subnational province units (e.g., Qatar). For the majority of model training and evaluation we use provincial HDI data from 2019, the same year as the MOSAIKS image features. We use provincial HDI data from 2012-2021 for mean-anchoring the global fine-resolution HDI estimates for that time period.

2.3 Municipality-level HDI

We compare our municipal HDI estimates with census-derived municipal estimates for HDI, where these data exist.

Indonesia Time series estimates of HDI at the municipality level are made publicly available by Badan Pusat Statistik (BPS), the statistics agency of Indonesia (*63*). We use estimates from 2019, which makes these, to our knowledge, the only available municipal HDI data that come from the same year as our satellite imagery.

Brazil The Human Development Report Office of the United Nations Development Programme has derived HDI data at the municipality level for Brazil, using census data (*35*). The most recent year these data are available is 2010, which we use in this analysis. The census-derived data have a different mean than the 2019 HDI data that we use elsewhere in this analysis; though, this has no influence on the within-country or within-province evaluation metrics (Figure 2 C,D) because predictions and observations are demeaned at the country and province level, respectively, before the metrics are calculated.

Mexico Census derived estimates of HDI are also available in Mexico for the year 2010, constructed by ref. (*9*). As with Brazil, these data have a different mean than the 2019 HDI data; though, again, that has no influence on the within-country or within-province evaluation metrics.

When certain components of HDI are not directly available at the municipal level, producers of these datasets use close proxies. For example the BPS calculates municipal HDI data in Indonesia using data on real expenditure rather than GNIpc (*34*). And when calculating municipal HDI in Mexico, ref. (*9*) use the child survival rate rather than life expectancy at birth for the health index and an asset index rather than GNIpc for the standard of living index. Likewise, provincial HDI estimates are also constructed using proxies when direct measurements of HDI components are not available (*7*). Discrepancies in how HDI is calculated at the municipal and provincial levels may contribute to differences between our satellite-based HDI estimates and these survey-based HDI estimates.

2.4 IWI Data

IWI data also come from GDL. These data are publicly available at the country and province levels and we use these data for 2019. GDL also provided us IWI data at the DHS cluster level, which are not publicly available. We use cluster-level IWI estimates from 2012 through 2019 in this analysis. We drop observations that do not overlap a parent province polygon and for which no imagery is available. This results in 51,996 DHS cluster observations from 41 countries.

We match all label data to time-constant satellite image features from 2019. Because these features are not contemporaneous with all labels, our results present a conservative estimate of the ability of SIML to measure HDI globally. Given that HDI variation is substantially larger over space than time, however, perfectly contemporaneous measures would likely improve performance only modestly (Supplementary Information S3).

3 Creation of features

3.1 MOSAIKS features

We create daytime image features using Planet's Surface Reflectance Basemaps product from 2019, which has a pixel resolution of 4.77m x 4.77m at the equator. These quarterly mosaics are processed by Planet to minimize cloud cover, balance color across seasons, and remove seams from images (*10*, *11*). We use data from quarter 3 because it corresponds with less ice coverage in the northern hemisphere and less cloud cover in the tropics. We follow the methods described in Rolf et al. (2021) to generate a set of 4000 task agnostic daytime image features using random convolutional features (*10*, *12*). Two thousand of these features use a patch size of 4 x 4 x 3 pixels, and the other two thousand use a patch size of 6 x 6 x 3 pixels. The third dimension of the patch size refers to the number of color bands (i.e., red,

green, and blue) that are available in Planet imagery. We selected these patch sizes because they maximized performance across three non-HDI prediction tasks: predicting nightlight intensity, road length, and forest cover at the global level. We tested patch sizes ranging from 3 x 3 x 3 to 10 x 10 x 3 and found that using a combination of two different patch sizes (with 2,000 patches each) outperformed using a single patch size (with 4,000 patches) across all three tasks.

We create features for all land tiles with available imagery on a global 0.01° x 0.01° equal-angle grid, amounting to ≈ 151 million feature vectors in total (*10*). Features become sparse above 60° latitude, due to a lack of available imagery. Figure 1C shows individual images spanning 0.01° x 0.01°, along with their corresponding MOSAIKS feature values.

We create polygon-level feature vectors by averaging values across the feature tiles associated with each polygon. Each administrative polygon is represented by a single vector of 4000 daytime image features. We assign each feature tile to the administrative polygon that contains its centroid. For small municipal and DHS polygons that do not contain any tile centroids, we represent the polygon by the nearest feature tile. When averaging, we weight by population using data from the Global Human Settlement Layer (GHS-POP) (*46*). We use the GHS-POP data product for the year 2020 at 30 arcsecond (0.008°) resolution.

3.2 Nighttime light features

We create non-linear NL features from the Visible Infrared Imaging Radiometer Suite (VIIRS) average masked data product (*32*). We use the V2.1 annual composite from the year 2019. The data has global coverage at 15 arcsecond (0.004°) resolution. Radiance units are expressed as $nW/cm^2/sr$. We assign the small number of pixels with negative radiance values to have a value of zero.

We create features that flexibly characterize the distribution of the NL data using indicator variables that represent whether the radiance value of each NL pixel falls into each of 21 bins. The first bin represents radiance values of zero. The next 20 bins represent radiance values that fall into evenly spaced quantiles of the global distribution of positive NL values. Analogous to aggregating the daytime imagery features to the polygon level, we calculate the population-weighted average value for each of the 21 bins for a given polygon. These polygon-level NL features denote the fraction of each polygon's population that is covered by NL values represented by each of the 21 bins. This approach allows NL to associate non-linearly with the outcome variables in our linear models.

We compute average nighttime light features for polygons by weighting by population in our main analysis. Thus, changes in VIIRS grid cell sizes caused by latitude do not affect

these feature values (*64*), since total population in each cell is measured independently of pixel area. We also verify that population-weighted features, which should have higher performance than area-weighted features because HDI is measured on a per-person basis (not per-area) do indeed have higher performance. For within-country provincial performance, $R^2 = 0.52$ using population-weighted features, and $R^2 = 0.24$ using area-weighted features, which were constructed accounting for the change in the VIIRS grid cell size with latitude.

Use of the "average masked" VIIRS data product should generally have background, biomass burning, and aurora radiance removed (*33*). To the extent there is still unmasked flaring in the NL data, we anticipate that the use of population-weighting when constructing the NL features minimizes this bias, as oil producing areas are typically sparsely populated.

4 Analysis

4.1 General model specification

All models are trained at either the country or province level and use either the 4000 MOSAIKS daytime imagery features, the 21 NL features, or both. We train models using a five-fold cross-validation procedure with basic ridge hyper-parameter tuning. Data are split by country during cross-validation to account for spatial autocorrelation and to ensure that the model is predicting provincial outcomes when no observations from within the same country have been observed. We apply a clipping procedure that restricts model predictions to the minimum and maximum value observed in the training data. Hyper-parameter tuning is done with this clipping procedure. We allow for a different hyper-parameter between the MOSAIKS and NL feature sets, though this has only a minor impact on our results.

The general linearized model, representing Eq. 3, that we implement is

$$Y = \beta_0 + \boldsymbol{\beta}_1 \, \mathbf{X}_{MOSAIKS} + \boldsymbol{\beta}_2 \, \mathbf{X}_{NL} + \epsilon \tag{4}$$

Where Y is used to refer to the HDI, IWI, or NL labels interchangeably. We use this same model but predict each outcome separately. $\mathbf{X}_{MOSAIKS}$ is the matrix of daytime MOSAIKS features and $\mathbf{X}_{NL}$ is the matrix of nightlight features. We learn β_0, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ using ridge regression, following Rolf et al. (*12*). When predicting the NL outcome, we always exclude the $\mathbf{X}_{NL}$ feature matrix. For each outcome, we report performance for models trained at the country, province, and within-country levels (Table S1). Model training is further detailed in Supplementary Information S6.1.

4.2 Performance metrics

For each model specification, we report two metrics, both of which are used in the literature. The *coefficient of determination* (R^2), used to evaluate related models by Chi *et al.* (*17*) and others, describes the accuracy of the raw model predictions and is a direct measure of model skill. The *square of the correlation coefficient* (ρ^2), used by Jean *et al.* (*20*) and others, scores performance after allowing model predictions to be linearly re-scaled before they are compared to observed values. We calculate both of these metrics when evaluating the full variation in labels and when decomposing the variation in labels into components that are visible within-countries or within-provinces (in contrast to between-countries and between-provinces).

Full variation performance The "full variation" performance metrics describe how well we estimate subnational HDI globally when we use all information available to us. To calculate full variation performance, we calculate ρ^2 and R^2 on the predicted and observed values of subnational HDI directly. This evaluates the ability of model predictions to capture the total variation in the observed values – i.e. variation across countries, across provinces within countries, and across municipalities or DHS clusters within provinces. Because most of the variation in HDI and other outcomes is between countries (Figure 2A-B), a large portion of the model's full variation performance comes directly from the mean-anchoring procedure (when it is used). Thus, the full variation performance metrics do not precisely evaluate the model's ability to predict local variation in isolation, and so they are not our preferred evaluation metric for understanding model performance within countries. Instead, we focus our analysis on within-country and, when applicable, within-province performance.

Within-country performance The "within-country" performance measures the amount of variation in the provincial deviations from the country mean that can be explained by the model. This metric evaluates the ability of the model to explain local variation in the outcome by removing large-scale variation in the outcome across countries in the demeaning step before predictions and observations are compared. To calculate within-country performance, we calculate ρ^2 and R^2 after demeaning predictions and observed values at the country level (i.e., after subtracting the predicted and observed country average value from each predicted and observed data point, respectively).

Within-province performance The "within-province" performance metric evaluates the ability of the model to explain hyper-local variation in the outcome, such as which DHS clusters within each province have higher or lower IWI. It does this by removing all between-

province variation in the predicted and observed values before they are compared. To calculate within-province performance, we calculate ρ^2 and R^2 after demeaning predictions and observed values at the province level.

4.3 Model evaluation

Evaluation of HDI at the same provincial resolution as model training When reporting model performance at the same resolution as training (Figure 2 top row, Table S1 upper section, and Table S3), we evaluate predictions from the validation folds of a five-fold spatial cross-validation procedure in which models are trained and evaluated on data from non-overlapping sets of countries. This enables more observations to be used when evaluating model performance. We also evaluate models on a held-out test set of countries that were not included when tuning the HDI model. Before analysis, we set aside $\approx 20\%$ of the provincial HDI data to be used as a final evaluation test set by randomly sampling 35 countries and their respective provinces. Evaluation on this test set was conducted after all hyper-parameter tuning and analysis decisions were made. We find that performance is not meaningfully different in the validation and tests sets, which indicates that the models evaluated on the validation folds did not over-fit to the data (Table S2).

Evaluation of HDI, IWI and NL at finer resolution than model training In the downscaling experiments, we evaluate performance using fine-resolution municipal or DHS cluster observations that were not used for model training or tuning. After tuning the model using cross-validation we retrain the model using the optimal hyper-parameters on all the province or country observations before predicting at downscaled resolution.

4.4 Mean-anchoring

In our primary within-country model, we anchor our estimates to country or province-level means depending on the experiment. Estimates from models trained on provincial or national observations in "levels" (Table S1) are never mean-anchored.

Anchoring to country means The procedure for anchoring to the country mean is illustrated in the top row of Figure 2, where we evaluate performance at the same resolution as model training. Our within-country model is trained to predict within-country anomalies, so in order to predict HDI in "levels" (Figure 2A), we add back the known country average HDI (Equation S2). This procedure enables us to calculate full variation performance (Figure 2A) but has no impact on the reported within-country performance (Figure 2B).

Anchoring to provincial means In the downscaling application, our goal is to produce the best possible estimates at fine resolution. Thus, when producing downscaled estimates of IWI and HDI, we anchor our predictions to the observed provincial value of the outcome (Equation S8). This re-centering procedure impacts the full variation performance and the within-country performance but does not impact the within-province performance (Table S1). Note that we anchor municipal NL estimates to country means, rather than provincial means, because we find that this improves the estimates (Supplementary Information S6.3). Mean anchoring and model evaluation are further detailed in Supplementary Information S6.1.

4.5 Producing downscaled estimates of HDI

To produce fine-resolution estimates of HDI, we take the local estimates of HDI that we predict using satellite imagery and mean-anchor them to province-year values from 2012-2021. This approach uses existing provincial values to explain provincial HDI over space and time, and then supplements that with information on local (i.e. within-province) variation in HDI predicted by the satellite imagery. This approach assume that the spatial pattern of HDI within provinces does not change over time; relaxing this assumption gives similar results (Supplementary Information S3). While in principle we could apply this approach to make estimates in years earlier than 2012, going farther back in time may make the assumption of a constant pattern of within-province HDI less strong. The year 2012 is the earliest for which we evaluated sub-provincial variation using IWI data.

Municipality-level HDI We use the within-country model specified in Equation S1a to estimate HDI at the municipality level, using a municipality (ADM2) shapefile from geoBoundaries (*65*). We anchor municipality estimates by centering predicted deviations on the observed province-level HDI value for each year, following the procedure for downscaled IWI predictions (Equation S8). We do not release HDI estimates for municipalities that cannot be linked to a parent province with a province-level HDI estimate from Smits and Permanyer (*7*) because there is not a known provincial value to anchor on.

Grid-level HDI To produce $0.1^\circ \times 0.1^\circ$ estimates of HDI, we similarly use the within-country model specified in Equation S1a. We make predictions using MOSAIKS features at $0.1^\circ \times 0.1^\circ$ resolution. This results in gridded estimates of HDI at approximately 10km^2 resolution. We mask out locations where humans are not believed to be settled based on GHS-POP (*46*) (keeping areas with population > 0) and then mean-anchor our tile estimates

such that the population-weighted average of the grid tiles within each province matches known provincial HDI values.

4.6 Targeting applications

In Figure 5 we evaluate how access to more granular HDI data improves the number of program recipients correctly targeted. We assume the goal is to provide a uniform transfer to those at or below the tenth percentile of the HDI distribution. Following Smythe and Blumenstock (2022) (*8*), the program is geographically targeted and all individuals within targeted regions receive the same transfer, a practice used to reduce administrative costs (*47*, *66*). If the administrator has access to only provincial HDI measures, then eligibility is determined at the province-level. Alternatively, using our municipal estimates, the administrator is able to target the program at the municipality level.

We evaluate performance using the census-derived municipality-level HDI measurements (*9*) — which are not used to train our model. These are the same census-derived measures used to evaluate performance in Figure 2C-D and discussed above (Methods 2.3). Because these estimates are municipality-level aggregates, we simulate HDI for *individuals* — i.e. the targets of the policy — by imposing additional assumptions about the distribution of HDI across individuals within a municipality. We assume HDI within each municipality has a truncated normal distribution (bounded between 0 and 1) that is centered around the census-derived municipality mean (Figure S6).

The degree to which municipal HDI measures improve targeting performance relative to provincial measures depends somewhat on the assumed dispersion of individual HDI values within municipalities (Methods 4.6). Assuming lower (higher) dispersion leads to larger (smaller) absolute — though similar proportional — gains (Figure 5E and Figure S6).

When evaluating performance using a receiver operating characteristic curve (ROC) curve, (Figure 5F) the aim is still to provide assistance to the 10% of individuals with the lowest HDI, but the fraction of the total population targeted is modified to examine how the true positive (individuals with low HDI correctly given aid) and false positive (individuals with HDI above the desired cutoff incorrectly given aid) rates change accordingly. Moving from left to right on the x-axis in Figure 5F implies a greater number of people receiving assistance. The area under the curve (AUC) shows the efficiency of the targeting. The AUC increases by 0.1 (+13% from 0.76 to 0.86) when municipal data are used instead of provincial data, indicating improved targeting performance when the budget constraint is varied.

As an additional robustness check, we repeat this policy targeting experiment using Indonesia municipality data (*63*) and observe a similar increase in AUC under the same

assumptions (11% from 0.62 to 0.69). This is shown in Figure S11.

Code availability Replication code is available at `github.com/Global-Policy-Lab/hdi_downscaling_mosaiks`.

Data availability All data used in this analysis, other than the DHS cluster-level IWI data from the Global Data Lab, is from free, publicly available sources. Details on how to access data for replication can be found at `github.com/Global-Policy-Lab/hdi_downscaling_mosaiks`. HDI estimates are available at `mosaiks.org/hdi`.

Funding This work was supported by a grant from the Human Development Report Office of the United Nations Development Programme. Additional support for this work comes from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752, the Harvard University Center for the Environment and Harvard University Data Science Initiative, and AI for Earth supported by Microsoft and National Geographic. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Acknowledgements We thank Tamma Carleton, Pedro Conceição, Esther Rolf, and seminar participants at The Workshop in Environmental Economics and Data Science, the WIDER Development Conference (Bogotá), IDinsight, and the American Geophysical Union Fall Meeting for their valuable feedback. We also thank Tamma Carleton for contributions to Fig. S1. We thank the Global Data Lab for sharing DHS cluster-level data on the International Wealth Index. We also thank Iñaki Permanyer and Arie Wahyu Wijayanto for helping us access census-derived municipality data in Mexico and Indonesia respectively. The project benefited from input and assistance from Jeanette Tseng, Jessica Katz, and Trinetta Chong.

References

1. P. Conceição, M. Kovacevic, T. Mukhopadhyay,
Human Development: A Perspective on Metrics, pp. 83–115, ISBN: 9780128190579.

2. I. Permanyer, J. Smits, *Population and Development Review*, ISSN: 17284457.

3. J. Klugman, F. Rodríguez, H. J. Choi,
The HDI 2010: New controversies, old critiques.
Journal of Economic Inequality **9**, 249–288, ISSN: 15691721 (2011).

4. *Technical notes - Human Development Report 2019*, 2019.

5. "Human Development Report 1990: Concept and Measurement of Human Development", tech. rep.
(United Nations Development Programme, New York, 1990).

6. H. Wolff, H. Chong, M. Auffhammer, Classification, Detection and Consequences of Data Error: Evidence from the Human Development Index.
Economic Journal **121**, 843–870, ISSN: 00130133 (2011).

7. J. Smits, I. Permanyer, The subnational human development database.
Scientific Data **6**, 1–15, ISSN: 20524463,
(`https://doi.org/10.1038/sdata.2019.38`) (2019).

8. I. S. Smythe, J. E. Blumenstock,
Geographic microtargeting of social assistance with high-resolution poverty maps.
Proceedings of the National Academy of Sciences **119**, e2120025119 (2022).

9. I. Permanyer,
Using Census Data to Explore the Spatial Distribution of Human Development.
World Development **46**, 1–13, ISSN: 0305-750X (June 2013).

10. T. Carleton *et al.*,
Multi-Task Observation Using Satellite Imagery and Kitchen Sinks (MOSAIKS) API,
`https://api.mosaiks.org`, version 1.0, 2022.

11. Planet Team, *Planet Application Program Interface: In Space for Life on Earth*,
Planet, 2017, (`https://api.planet.com`).

12. E. Rolf *et al.*, A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications 2021 12:1* **12**, 1–11, ISSN: 2041-1723
(July 2021).

13. P. M. Atkinson, Downscaling in remote sensing.
International Journal of Applied Earth Observation and Geoinformation **22**, Spatial Statistics for Mapping the Environment, 106–114, ISSN: 1569-8432 (2013).

14. A. N. Rose, E. Bright, The LandScan global population distribution project: Current state of the art and prospective innovation.
Population Assiociation of America 2014 Annual Meeting, 1–21 (2014).

15. M. Burke, A. Driscoll, D. B. Lobell, S. Ermon,
Using satellite imagery to understand and promote sustainable development.
Science **371**, eabe8628 (2021).

16. C. Yeh *et al.*, Using publicly available satellite imagery and deep learning to understand economic well-being in Africa.
Nature Communications **11**, 1–11, ISSN: 20411723 (Dec. 2020).

17. G. Chi, H. Fang, S. Chatterjee, J. E. Blumenstock,
Microestimates of wealth for all low-and middle-income countries.
Proceedings of the National Academy of Sciences **119**, e2113658119 (2022).

18. A. Khachiyan *et al.*,
Geographic microtargeting of social assistance with high-resolution poverty maps.
American Economic Review: Insights **4**, 491–506 (2022).

19. R. Engstrom, J. Hersh, D. Newhouse, Poverty from space: Using high resolution satellite imagery for estimating economic well-being.
The World Bank Economic Review **36**, 382–412 (2022).

20. N. Jean *et al.*, Combining satellite imagery and machine learning to predict poverty.
Science **353**, 790–794, ISSN: 10959203 (Aug. 2016).

21. X. Zhang, J. Xu, S. Zhong, Z. Wang,
Assessing uneven regional development using nighttime light satellite data and machine learning methods: evidence from county-level improved HDI in China.
Land **13**, 1524 (2024).

22. R. Ramadhan, A. W. Wijayanto, presented at the Proceedings of The International Conference on Data Science and Official Statistics, vol. 2023, pp. 274–295.

23. E. Aiken, E. Rolf, J. Blumenstock,
Fairness and representation in satellite-based poverty maps: Evidence of urban-rural disparities and their impacts on downstream policy, 2023,
arXiv: `2305.01783 (cs.LG)`.

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

25. J. Proctor, T. Carleton, S. Sum, "Parameter Recovery Using Remotely Sensed Variables", tech. rep. (National Bureau of Economic Research, 2023).

26. E. Aiken, S. Bellue, D. Karlan, C. Udry, J. E. Blumenstock, Machine learning and phone data can improve targeting of humanitarian aid. *Nature* **603**, 864–870 (2022).

27. L. Y. Huang, S. M. Hsiang, M. Gonzalez-Navarro, "Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs", tech. rep. (National Bureau of Economic Research, 2021).

28. Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21 (2021).

29. N. Jean *et al.*, Tile2vec: Unsupervised representation learning for spatially distributed data. *Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 3967–3974 (2019).

30. S. Mohanasundaram, K. Kasiviswanathan, C. Purnanjali, I. P. Santikayasa, S. Singh, Downscaling Global Gridded Crop Yield Data Products and Crop Water Productivity Mapping Using Remote Sensing Derived Variables in the South Asia. *International Journal of Plant Production*, 1–16 (2022).

31. P. Rayner, M. Raupach, M. Paget, P. Peylin, E. Koffi, A new global gridded data set of CO2 emissions from fossil fuel combustion: Methodology and evaluation. *Journal of Geophysical Research: Atmospheres* **115** (2010).

32. *Annual VIIRS NL V2.1*, (`https://eogdata.mines.edu/products/vnl/`).

33. C. D. Elvidge, M. Zhizhin, T. Ghosh, F.-C. Hsu, J. Taneja, Annual Time Series of Global VIIRS Nighttime Lights Derived from Monthly Averages: 2012 to 2019. *Remote Sensing* **13**, ISSN: 2072-4292, (`https://www.mdpi.com/2072-4292/13/5/922`) (2021).

34. *Indeks Pembangunan Manusia 2023*, 2024, (2025; `https://www.bps.go.id/en/publication/2024/05/13/8f77e73a66a6f484c655985a/indeks-pembangunan-manusia-2023.html`).

35. "Desenvolvimento Humano Para Além Das Médias", tech. rep. (United Nations Development Programme & Instituto de Pesquisa Ecônomica Aplicada, Brasília, 2017), (2025; https://portalantigo.ipea.gov.br/portal/index.php?option=com_content&view=article&id=30025).

36. J. Smits, R. Steendijk, The International Wealth Index (IWI). *Social Indicators Research 2014 122:1* **122**, 65–85, ISSN: 1573-0921 (July 2014).

37. X. Chen, W. D. Nordhaus, Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences* **108**, 8589–8594, ISSN: 0027-8424 (May 2011).

38. J. V. Henderson, A. Storeygard, D. N. Weil, Measuring Economic Growth from Outer Space. *American Economic Review* **102**, 994–1028 (Apr. 2012).

39. A. Bruederle, R. Hodler, Nighttime lights as a proxy for human development at the local level. *PLOS ONE* **13**, 1–22 (Sept. 2018).

40. R. Bluhm, G. C. McCord, What Can We Learn from Nighttime Lights for Small Geographies? Measurement Errors and Heterogeneous Elasticities. *Remote Sensing* **14**, ISSN: 2072-4292 (2022).

41. C. D. Elvidge, K. E. Baugh, S. J. Anderson, P. C. Sutton, T. Ghosh, The Night Light Development Index (NLDI): a spatially explicit measure of human development from satellite data. *Social Geography* **7**, 23–35 (2012).

42. C. D. Elvidge *et al.*, A global poverty map derived from satellite data. *Computers & Geosciences* **35**, 1652–1660, ISSN: 0098-3004 (2009).

43. P. Sutton, C. Elvidge, G. Tilottama, Estimation of Gross Domestic Product at Sub-National Scales Using Nighttime Satellite Imagery. *International Journal of Ecological Economics & Statistics* **8** (Jan. 2007).

44. I. McCallum *et al.*, Estimating global economic well-being with unlit settlements. *Nature Communications 2022 13:1* **13**, 1–8, ISSN: 2041-1723 (May 2022).

45. *UNDP Human Development Data Center*, 2019, (http://hdr.undp.org/en/data).

46. *Global Human Settlement Layer - Population (GHS POP E2020)*, (https://ghsl.jrc.ec.europa.eu/download.php?ds=pop).

47. C. Elbers, T. Fujii, P. Lanjouw, B. Özler, W. Yin,
Poverty alleviation through geographic targeting: How much does disaggregation help?
Journal of Development Economics **83**, 198–213 (2007).

48. M. Ravallion,
Poverty alleviation through regional targeting: A case study for Indonesia.
The Economics of Rural Organization: Theory, Practice and Policy, 373–77 (1993).

49. D. P. Coady, The welfare returns to finer targeting: The case of the PROGRESA program in Mexico. *International tax and public finance* **13**, 217–239 (2006).

50. I. Smythe, J. Blumenstock,
Geographic Micro-Targeting of Social Assistance with High-Resolution Poverty Maps.
Proceedings of ACM Conference (Conference'17) **1** (2021).

51. C. Yeh *et al.*, SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning. arXiv: `2111.04724` (Nov. 2021).

52. N. Pokhriyal, D. C. Jacques,
Combining disparate data sources for improved poverty prediction and mapping.
Proceedings of the National Academy of Sciences **114**, E9783–E9792 (2017).

53. P. Corral, I. Molina, A. Cojocaru, S. Segovia,
Guidelines to small area estimation for poverty mapping
(World Bank Washington, 2022).

54. D. Newhouse, Small Area Estimation of Poverty and Wealth Using Geospatial Data: What Have We Learned So Far? *Calcutta Statistical Association Bulletin* **76**, 7–32 (2024).

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

56. R. Van Der Weide, B. Blankespoor, C. Elbers, P. Lanjouw, How accurate is a poverty map based on remote sensing data? An application to Malawi.
Journal of Development Economics **171**, 103352 (2024).

57. O. Hall, M. Ohlsson, T. Rögnvaldsson, A review of explainable AI in the satellite data, deep machine learning, and human poverty domain.
Patterns **3**, 100600, ISSN: 2666-3899 (2022).

58. C. Coglianese, D. Lehr,
Regulating by Robot: Administrative Decision Making in the Machine-Learning Era.
Georgetown Law Journal **105** (2017).

59. J. L. Abitbol, M. Karsai, Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nature Machine Intelligence* **2**, 684–692, ISSN: 2522-5839 (2020).

60. K. Ayush, B. Uzkent, M. Burke, D. Lobell, S. Ermon, presented at the Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, ed. by C. Bessiere, Special track on AI for CompSust and Human well-being, pp. 4410–4416.

61. A. Rahimi, B. Recht, Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems* **21** (2008).

62. *Subnational-HDI Database V7.0*, (`https://globaldatalab.org/shdi/download_files/`).

63. *BPS-Statitistics Indonesia - Data Query*, (2025; `https://www.bps.go.id/en/query-builder`).

64. C. D. Elvidge, T. Ghosh, F.-C. Hsu, M. Zhizhin, M. Bazilian, The dimming of lights in China during the COVID-19 pandemic. *Remote Sensing* **12**, 2851 (2020).

65. D. Runfola *et al.*, geoBoundaries: A global database of political administrative boundaries (CGAZ 3.0.0). *PLOS ONE* **15**, e0231866, ISSN: 1932-6203 (Apr. 2020).

66. R. Hanna, B. A. Olken, Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives* **32**, 201–26 (2018).

67. *Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A (SR)*, (2025; `https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED`).

68. A. V. Norström *et al.*, Principles for knowledge co-production in sustainability research. *Nature sustainability* **3**, 182–190 (2020).

69. A. Kliskey *et al.*, Building trust, building futures: Knowledge co-production as relationship, design, and process in transdisciplinary science. *Frontiers in Environmental Science* **11**, 137 (2023).

70. *Microsoft Bing Maps - Global ML Building Footprints*, (2024; `https://github.com/microsoft/GlobalMLBuildingFootprints/tree/main?tab=readme-ov-file`).

71. *Open Street Maps API*, (2021; https://wiki.openstreetmap.org/wiki/API_v0.6).

72. A. Shaver, D. B. Carter, T. W. Shawa,
Terrain ruggedness and land cover: Improved data for most research designs.
Conflict Management and Peace Science **36**, 191–218 (2019).

73. M. C. Hansen *et al.*, High-resolution global maps of 21st-century forest cover change.
Science (New York, N.Y.) **342**, 850–3, ISSN: 1095-9203 (Nov. 2013).

74. N. Ramankutty, A. T. Evan, C. Monfreda, J. A. Foley, Farming the planet: 1.
Geographic distribution of global agricultural lands in the year 2000.
Global Biogeochemical Cycles **22** (2008).

75. *EarthStat*, (2021; http://www.earthstat.org/).

76. *CPC Global Unified Temperature*,
(2024; https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html).

77. *NASA Global Precipitaion Measurement. IMERG Grand Average Precipitation Climatology*,
(2024; https://gpm.nasa.gov/data/imerg/precipitation-climatology).

Supplementary Information

	Predicted at province level (n=1,381)					
	Full variation performance		*Within-country performance*			
	ρ^2	R^2	ρ^2	R^2		
	(1)	(2)	(3)	(4)		
HDI trained at:						
Within-country (n=1,363)	0.96	0.96	0.52	0.52		
Province level (n=1,381)	0.83	0.83	0.44	0.24		
Country level (n=145)	0.74	0.74	0.28	< 0		

	Predicted at municipality level in Indonesia (n=505)			
	Within-country performance		*Within-province performance*	
	ρ^2	R^2	ρ^2	R^2
	(3)	(4)	(5)	(6)
HDI trained at:				
Within-country (n=1,363)	0.62	0.61	0.53	0.53
Province level (n=1,381)	0.5	0.37	0.51	0.37
Country level (n=145)	0.35	< 0	0.36	< 0

	Predicted at municipality level in Brazil (n=5,584)			
	Within-country performance		*Within-province performance*	
	ρ^2	R^2	ρ^2	R^2
	(3)	(4)	(5)	(6)
HDI trained at:				
Within-country (n=1,363)	0.48	0.48	0.33	0.20
Province level (n=1,381)	0.46	0.36	0.31	< 0
Country level (n=145)	0.29	< 0	0.18	< 0

	Predicted at municipality level in Mexico (n=2,457)			
	Within-country performance		*Within-province performance*	
	ρ^2	R^2	ρ^2	R^2
	(3)	(4)	(5)	(6)
HDI trained at:				
Within-country (n=1,363)	0.5	0.45	0.31	0.29
Province level (n=1,381)	0.31	0.31	0.26	0.23
Country level (n=145)	0.16	< 0	0.12	< 0

	Predicted at DHS cluster level (n=51,996)					
	Full variation performance		*Within-country performance*		*Within-province performance*	
	ρ^2	R^2	ρ^2	R^2	ρ^2	R^2
	(1)	(2)	(3)	(4)	(5)	(6)
IWI trained at:						
Within-country (n=862)	0.71	0.67	0.56	0.56	0.36	0.36
Province level (n=862)	0.44	0.28	0.21	0.08	0.24	0.1
Country level (n=85)	0.31	< 0	0.13	< 0	0.13	< 0

	Predicted at municipality level (n=62,536)					
	Full variation performance		*Within-country performance*		*Within-province performance*	
	ρ^2	R^2	ρ^2	R^2	ρ^2	R^2
	(1)	(2)	(3)	(4)	(5)	(6)
NL trained only on MOSAIKS at:						
Within-country (n=2,852)	0.73	0.7	0.64	0.63	0.61	0.57
Province level (n=2,852)	0.65	0.61	0.57	0.45	0.53	0.35
Country level (n=170)	0.44	0.3	0.36	0.08	0.31	< 0

Table S1: Performance for models trained to predict HDI, IWI, and population-weighted nightlight luminosity (NL). Models for HDI and IWI use a combination of MOSAIKS and population-weighted NL features. We show performance evaluated at the province level for HDI and evaluate downscaled performance for HDI, IWI, and NL. Performance scatters from the within-country models are shown in Figure 2. All predictions are made for the year 2019.

		Predicted at province level (n=378)			
		Full variation performance		*Within-country performance*	
		ρ^2	R^2	ρ^2	R^2
		(1)	(2)	(3)	(4)
HDI trained at:	***Features***				
Within-country	**MOSAIKS+NL**	0.97	0.97	0.43	0.42
Province level	**MOSAIKS+NL**	0.87	0.87	0.4	0.09
Country level	**MOSAIKS+NL**	0.79	0.79	0.29	< 0

Table S2: This is similar to the upper portion of Table S1 except that here we have evaluated on a 35 country ($\approx$ 20%) test set that was not used during model tuning. The modest reported differences in the validation-set and test-set performances when evaluating within-country performance could be due to either noise from the small sample size of the test set, or to overfitting. We test this and find that noise from the small test set is likely to be the explanation. The average within-country test-set performance across 30 random 80% validation-set and 20% test-set splits, using the same training and evaluation procedure, is very close to that of the original validation set: $R^2 = 0.51$ for MOSAIKS + NL.

		Predicted at province level (n=1,381)			
		Full variation performance		*Within-country performance*	
		ρ^2	R^2	ρ^2	R^2
		(1)	(2)	(3)	(4)
HDI trained at:	***Features***				
Within-country (n=1,363)	**MOSAIKS+NL**	0.96	0.96	0.52	0.52
	MOSAIKS	0.95	0.95	0.42	0.42
	NL	0.95	0.95	0.45	0.45
Province level (n=1,381)	**MOSAIKS+NL**	0.83	0.83	0.44	0.24
	MOSAIKS	0.76	0.75	0.0.31	< 0
	NL	0.60	0.60	0.44	< 0
Country level (n=145)	**MOSAIKS+NL**	0.74	0.74	0.28	< 0
	MOSAIKS	0.62	0.58	0.16	< 0
	NL	0.59	0.53	0.44	< 0
		Predicted at province level (n=1,381)			
		Full variation performance		*Within-country performance*	
		ρ^2	R^2	ρ^2	R^2
		(1)	(2)	(3)	(4)
Life expectancy trained at:	***Features***				
Within-country (n=1,363)	**MOSAIKS+NL**	0.92	0.92	0.05	0.05
	MOSAIKS	0.92	0.92	0.01	0.01
	NL	0.92	0.92	0.05	0.05
Province level (n=1,381)	**MOSAIKS+NL**	0.69	0.69	0.03	< 0
	MOSAIKS	0.66	0.66	0.02	< 0
	NL	0.43	0.43	0.06	< 0
Country level (n=145)	**MOSAIKS+NL**	0.6	0.57	0.02	< 0
	MOSAIKS	0.57	0.53	0.02	< 0
	NL	0.42	0.32	0.06	< 0
		Predicted at province level (n=1,381)			
		Full variation performance		*Within-country performance*	
		ρ^2	R^2	ρ^2	R^2
		(1)	(2)	(3)	(4)
Mean years schooling trained at:	***Features***				
Within-country (n=1,363)	**MOSAIKS+NL**	0.93	0.93	0.51	0.51
	MOSAIKS	0.91	0.91	0.41	0.41
	NL	0.92	0.92	0.45	0.45
Province level (n=1,381)	**MOSAIKS+NL**	0.75	0.75	0.49	0.43
	MOSAIKS	0.7	0.7	0.41	0.31
	NL	0.56	0.56	0.44	0.26
Country level (n=145)	**MOSAIKS+NL**	0.73	0.72	0.46	0.3
	MOSAIKS	0.6	0.6	0.24	< 0
	NL	0.56	0.53	0.42	< 0
		Predicted at province level (n=1,381)			
		Full variation performance		*Within-country performance*	
		ρ^2	R^2	ρ^2	R^2
		(1)	(2)	(3)	(4)
Expected years schooling trained at:	***Features***				
Within-country (n=1,363)	**MOSAIKS+NL**	0.9	0.9	0.28	0.27
	MOSAIKS	0.89	0.89	0.27	0.26
	NL	0.88	0.88	0.15	0.15
Province level (n=1,381)	**MOSAIKS+NL**	0.56	0.56	0.21	0.09
	MOSAIKS	0.5	0.5	0.2	0.06
	NL	0.39	0.39	0.15	< 0
Country level (n=145)	**MOSAIKS+NL**	0.54	0.53	0.2	< 0
	MOSAIKS	0.44	0.41	0.07	< 0
	NL	0.38	0.35	0.14	< 0
		Predicted at province level (n=1,381)			
		Full variation performance		*Within-country performance*	
		ρ^2	R^2	ρ^2	R^2
		(1)	(2)	(3)	(4)
GNIpc trained at:	***Features***				
Within-country (n=1,363)	**MOSAIKS+NL**	0.97	0.97	0.56	0.56
	MOSAIKS	0.95	0.95	0.4	0.4
	NL	0.96	0.96	0.55	0.55
Province level (n=1,381)	**MOSAIKS+NL**	0.79	0.79	0.36	< 0
	MOSAIKS	0.68	0.68	0.25	< 0
	NL	0.59	0.59	0.52	< 0
Country level (n=145)	**MOSAIKS+NL**	0.56	< 0	0.23	< 0
	MOSAIKS	0.46	< 0	0.1	< 0
	NL	0.31	< 0	0.12	< 0

Table S3: Similar to the top section of Table S1 except that here we show performance for each HDI component and also show performance with different combinations of features.

	HDI	Life expectancy	Mean years schooling	Expected years schooling
Life expectancy	0.79			
Mean years schooling	0.84	0.52		
Expected years schooling	0.82	0.57	0.61	
GNIpc	0.62	0.46	0.5	0.44
Within-country	**HDI**	**Life expectancy**	**Mean years schooling**	**Expected years schooling**
Life expectancy	0.31			
Mean years schooling	0.82	0.12		
Expected years schooling	0.65	0.1	0.46	
GNIpc	0.17	0.03	0.1	0.07

Table S4: Individual components of HDI tend to be correlated. We report the squared Pearson's correlation coefficient (ρ^2) between HDI and its components at the province level. We also report the squared correlation coefficients after demeaning provincial observations by country. This ρ^2 metric used here is intended to be comparable to the metrics reported in Tables S1 and S3. Notably, within-country correlation between HDI and GNIpc is low, yet we are still able to predict those separate outcomes with considerable skill.

		Predicted at province level			
		Full variation performance		*Within-country performance*	
		ρ^2	R^2	ρ^2	R^2
		(1)	(2)	(3)	(4)
HDI trained at:	***Features***				
Within-country (n=1,363)	**NDVI+NDWI+NDBI**	0.92	0.92	0.03	0.02
	NDVI+NDWI+NDBI+MOSAIKS+NL	0.96	0.96	0.52	0.52
	MOSAIKS+NL (*for reference*)	0.96	0.96	0.52	0.52

Table S5: Performance for models trained with additional features at the provincial level. Specifically, we use Sentinel 2A imagery downloaded at approximately 500m resolution (*67*). Following, (*22*) we process Sentinel 2A imagery to calculate the Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Normalized Difference Built-up Index (NDBI). We then create population-weighted features in the same manner as done with the NL features (see Methods 3.2).

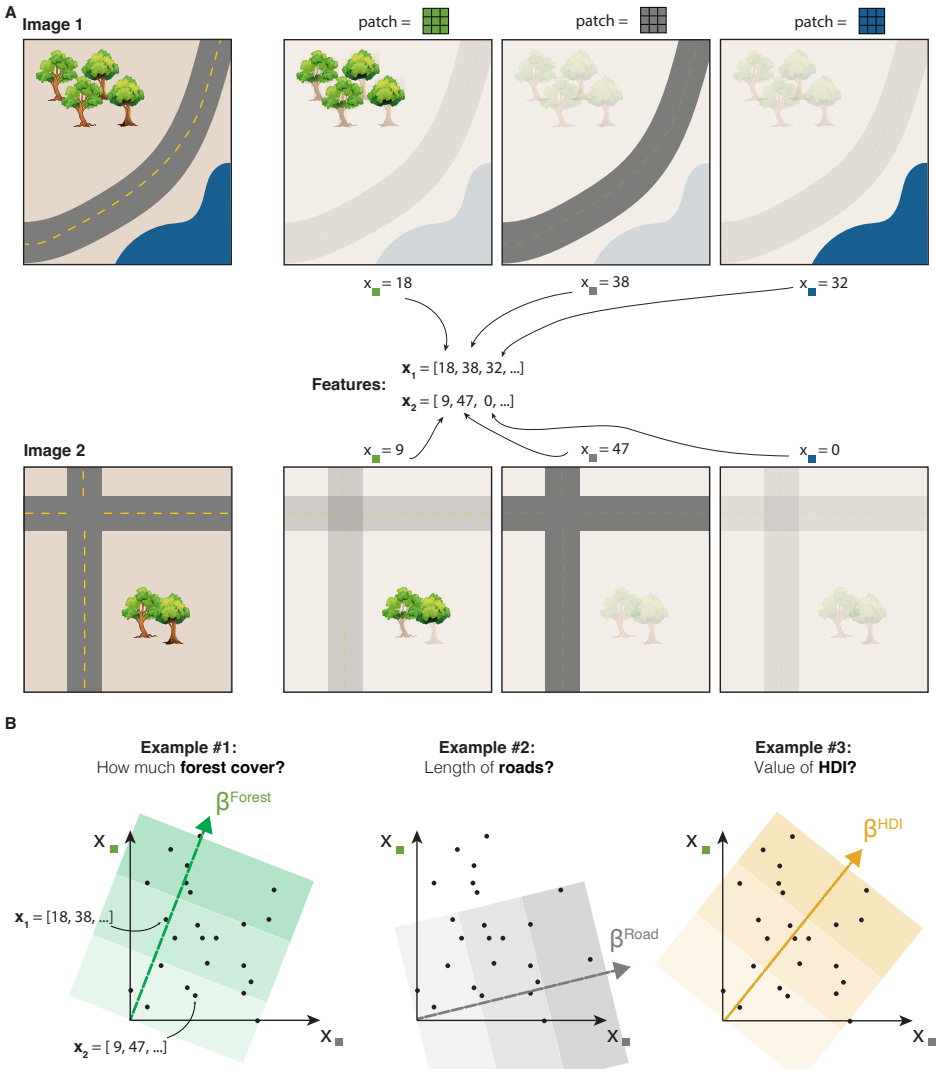

Figure S1: **An illustration of generating MOSAIKS features for two cartoon images, and using the features to predict HDI and other example outcomes.** (**A**) MOSAIKS random convolutional features capture the information within satellite imagery by measuring how similar each image is to a fixed set of small patches of imagery. Similarity is measured mathematically using a moving-window dot product, or "convolution." The parts of each image that are similar to each patch are shown, with greater similarity leading to larger feature values. The green patch, for example, is similar to the parts of the imagery containing green trees. Image 1 has a greater feature value for this patch because it has more trees than image 2. Collectively, MOSAIKS features capture information on the color and texture of the imagery, which represent the content of the imagery (e.g., trees, roads, and lakes). (**B**) Features associate differently with different outcomes: images that are more similar to the green patch tend to have higher forest cover, and images that are more similar to the grey patch tend to have more roads. Regressing HDI onto these features learns how higher or lower feature values associate with higher or lower HDI, and in turn, how to predict HDI using these features. Each dot in each scatter represents an image, and the arrow represents the direction in the feature space of increasing forest, roads, or HDI. The direction of the arrow is learned by the regression. For more details on MOSAIKS features and how they can be used to predict a broad range of outcomes see ref. (*12*).

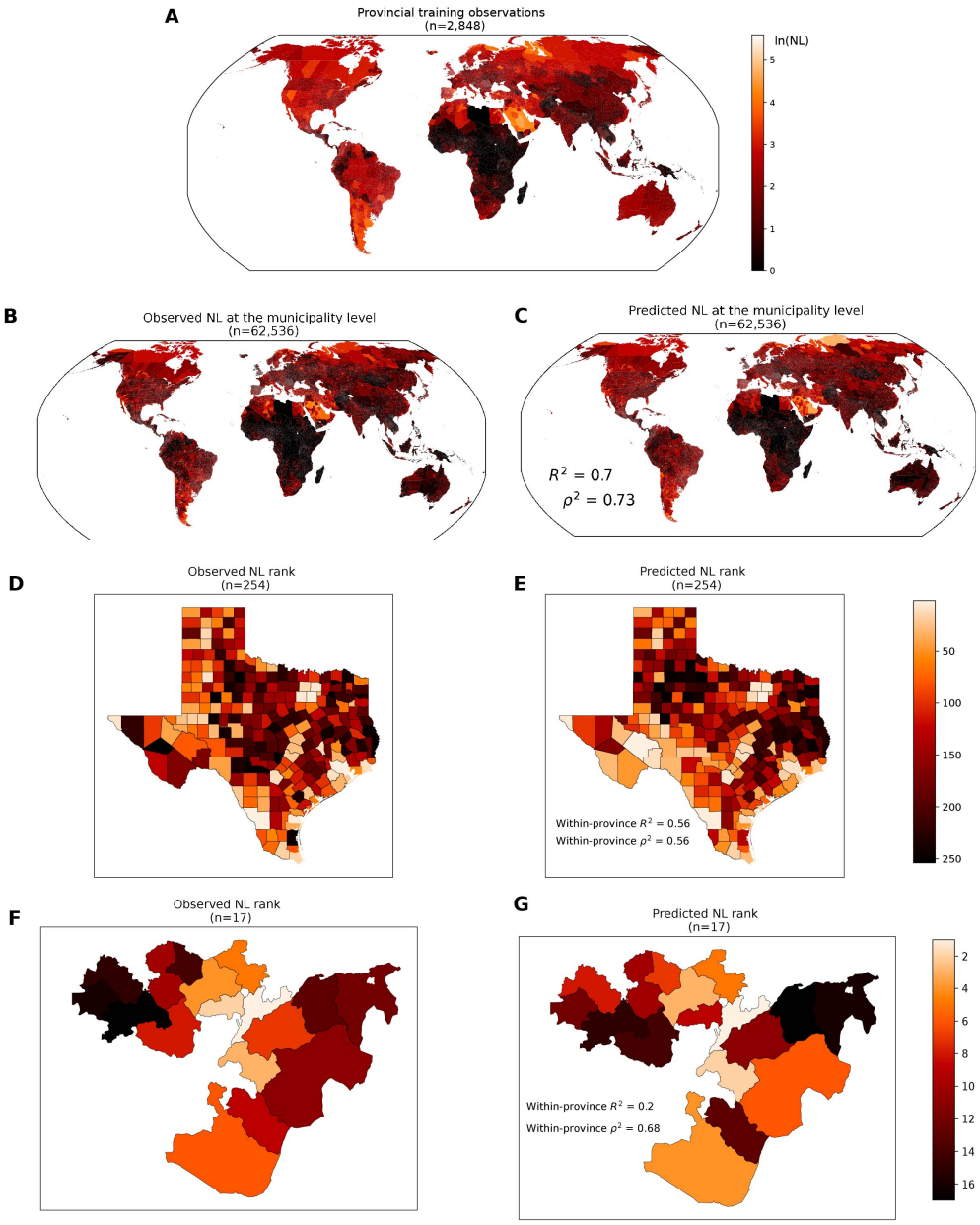

Figure S2: **A MOSAIKS model trained at the province level can effectively predict NL at the municipality level.** These maps show population-weighted NL luminosity that has been predicted using MOSAIKS. (**A**) Population-weighted NL averaged up to the provincial polygon for 2019. These are the data used to train the model. (**B**) True population-weighted NL at the municipality level. (**C**) Predicted population-weighted NL at the municipality level. (**D**) Municipalities ranked by luminosity within Texas, a single province in the United States. (**E**) Predicted nightlight luminosity rank within Texas. (**F**) Municipalities ranked by luminosity within Oromia, a single province in Ethiopia. (**G**) Predicted nightlight luminosity rank within Oromia. Panels D-G illustrate the downscaling efficacy of MOSAIKS. Each of these polygons (Texas and Oromia) represent a single training observation. All predictions come from a within-country model with predictions anchored to the country mean. Note that panels A-C use the same colorbar. See Table S1 for detailed performance metrics.

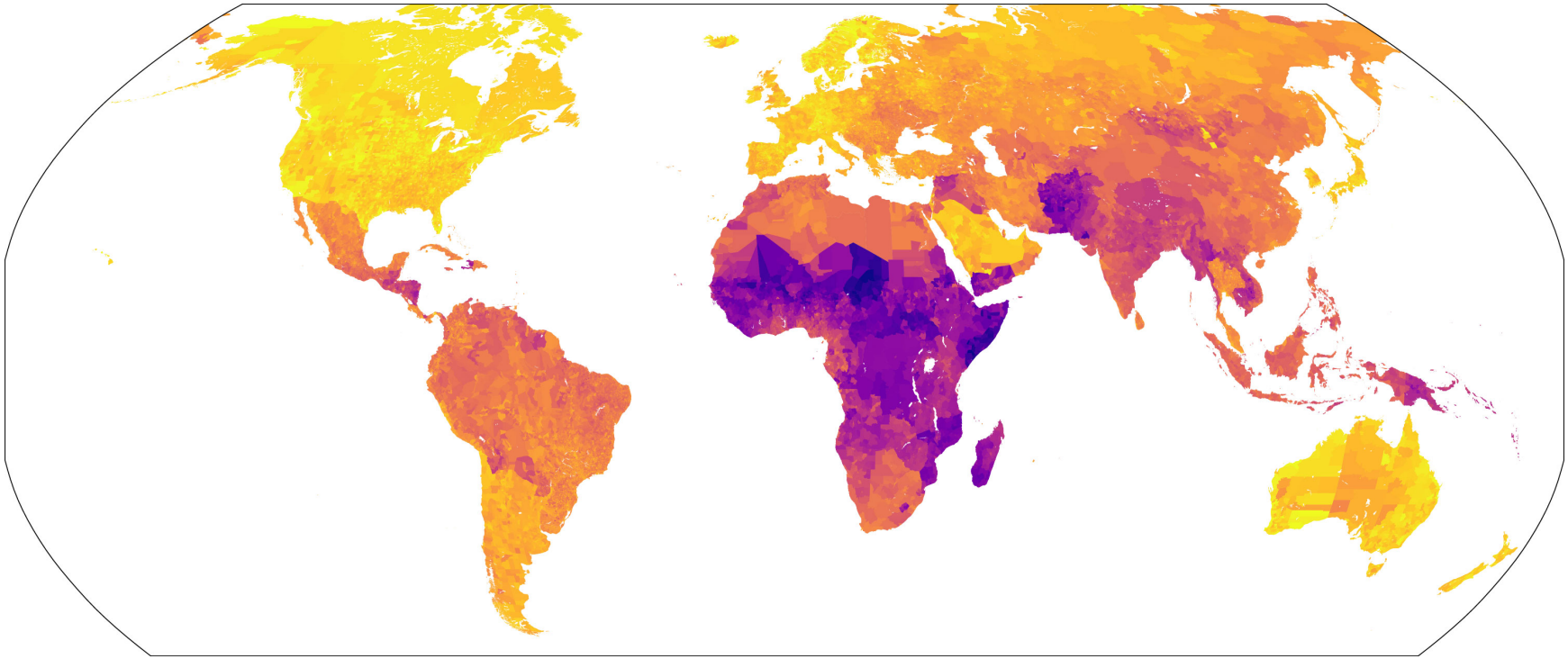

Figure S3: **Full page map of HDI estimates at the municipal level.** This is the same data as shown in Figure 3C.

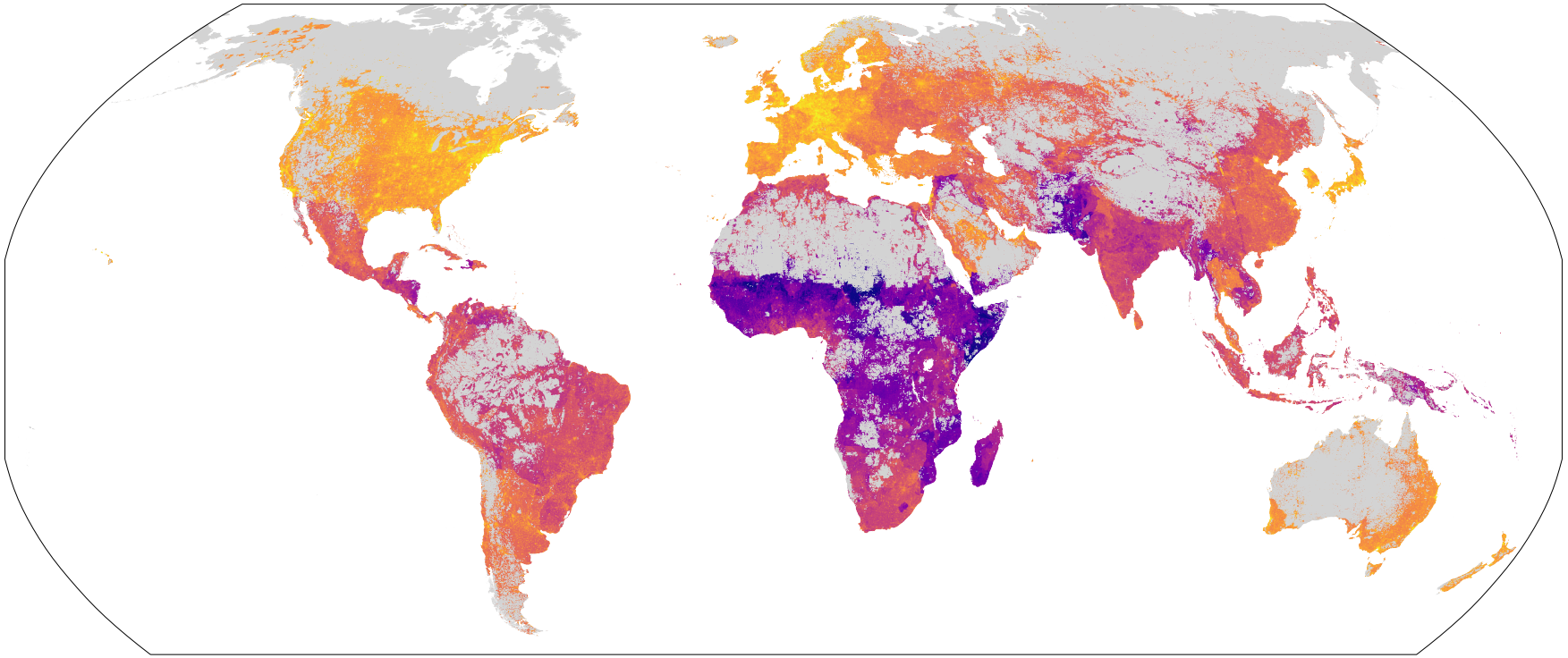

Figure S4: **Full page map of HDI estimates at the grid level.** This is the same data as shown in Figure 3D.

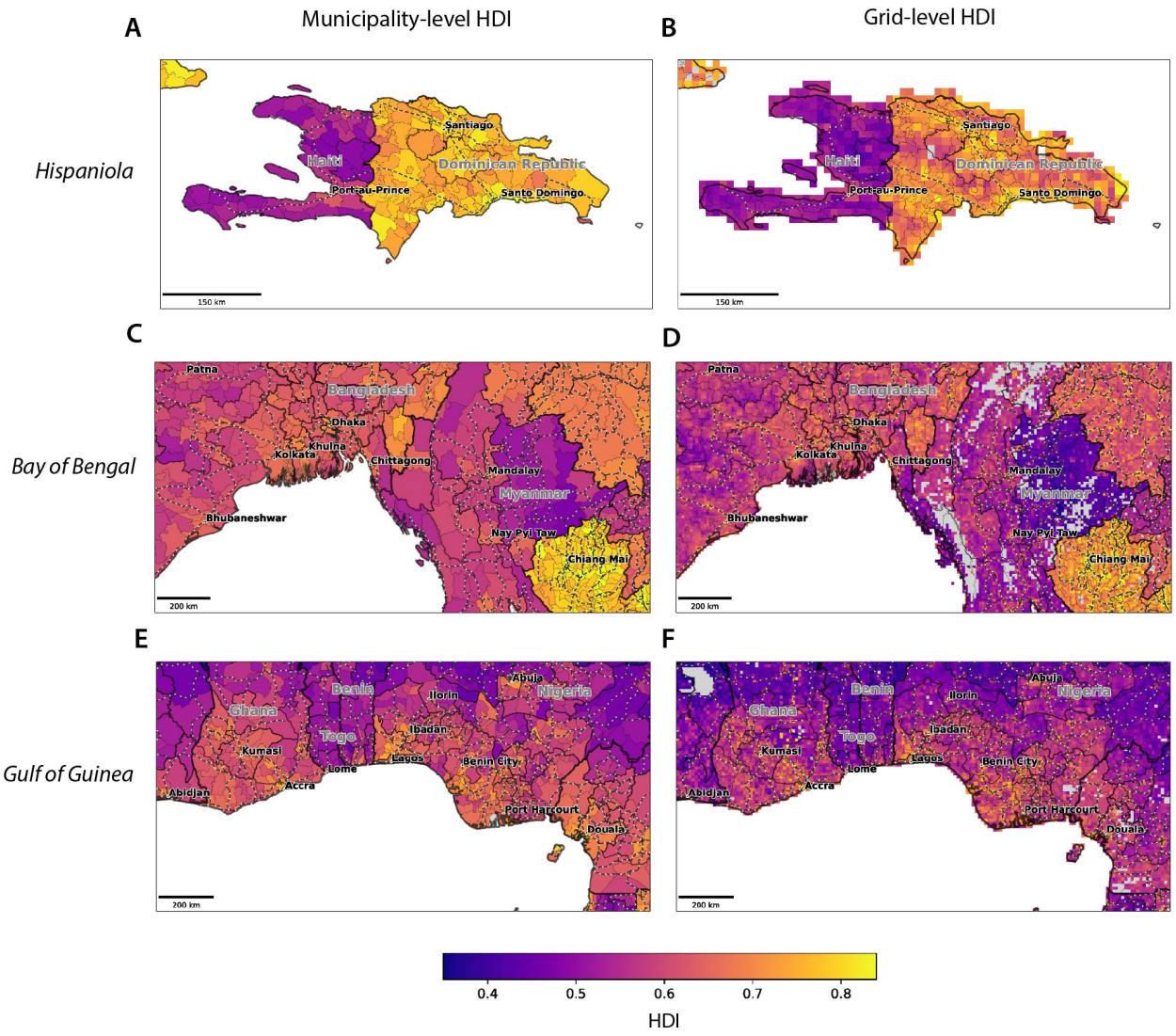

Figure S5: **Regional maps of HDI estimates at the municipal and grid levels.** (**A-B**) HDI estimates on Hispaniola (**C-D**) HDI estimates around the Bay of Bengal (**E-F**) HDI estimates around the Gulf of Guinea. All panels show country, province, and municipality borders as solid lines. Dashed lines show major roadways. Grey in the grid-level estimates indicates land area believed to be unsettled (*46*).

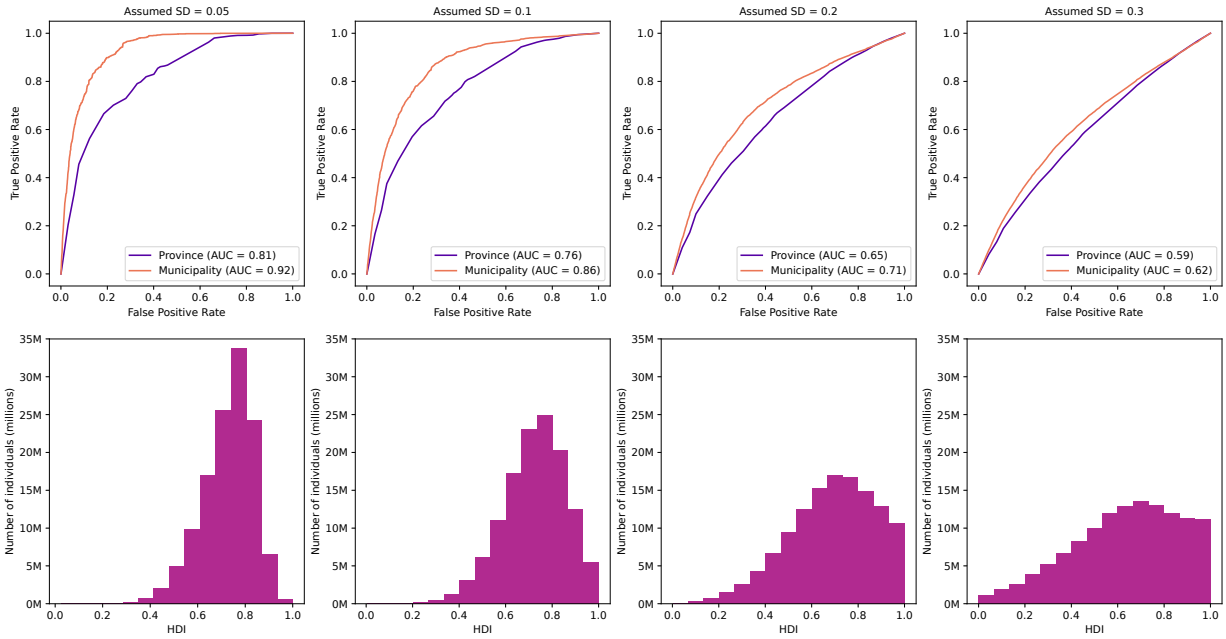

Figure S6: **The improvement in geographic targeting efficacy from using municipal values (ADM2) depends on the assumed variability of individual-level HDI within municipalities**. ROC curves as in Figure 5F for different assumed standard deviations (SD) of individual-level HDI within municipalities. Using municipal instead of provincial HDI estimates increases the AUC by 0.11 (+14% from 0.81 to 0.92) when the within-municipality HDI standard deviation is assumed to be 0.05 and by 0.06 (+9% from 0.65 to 0.71) when it is assumed to be 0.2. Histograms show the distribution of simulated individual-level HDI for each assumed SD, using a truncated normal distribution centered on the municipal values for Mexico calculated by Permanyer (*9*).

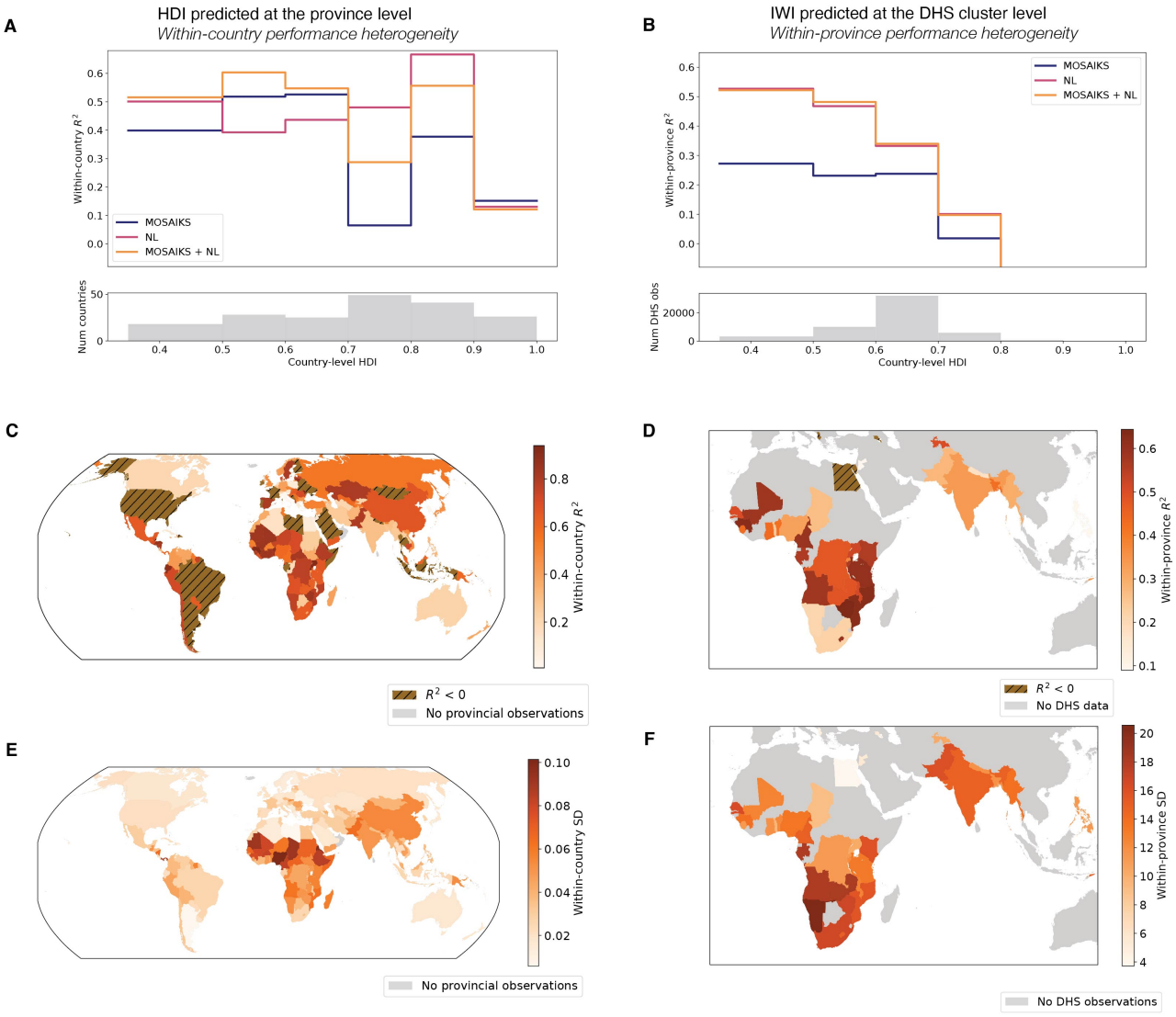

Figure S7: **Heterogeneity of HDI and IWI predictions.** On the left we show heterogeneity in HDI performance evaluated at the province level. On the right, we show heterogeneity in IWI performance evaluated at the DHS cluster level. (**A**) HDI performance as a function of parent country HDI. (**B**) IWI performance as a function of parent country HDI. (**C**) Mapped performance of HDI within-countries (within-country MOSAIKS + NL model). (**D**) Mapped performance of IWI within-provinces (within-country MOSAIKS + NL model). (**E**) Standard deviation of provincial HDI by country (**F**) Standard deviation of DHS cluster-level IWI within-provinces by country.

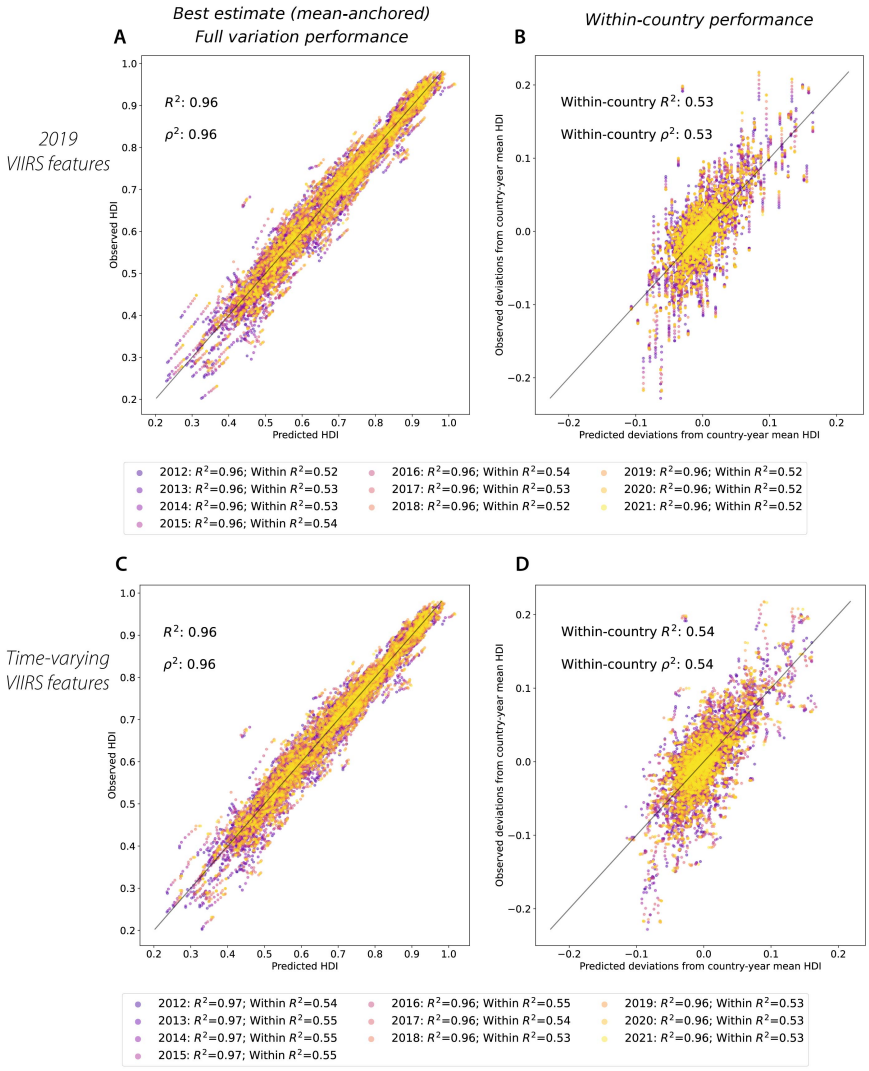

Figure S8: **Prediction of provincial HDI from 2012 to 2021 using satellite imagery. (A-B)** This figure is a replication of Figure 2A-B evaluated using a panel of provincial HDI from 2012-2021 and time static VIIRS features from 2019. Full variation performance is shown on the left and within-country performance is shown on the right. Each point represents a provincial-year observation and points are colored by year. Predicted provincial values are mean-anchored to the known country-year value. **(C-D)** This is a a similar replication of Figure 2A-B trained and evaluated using annually varying VIIRS features.

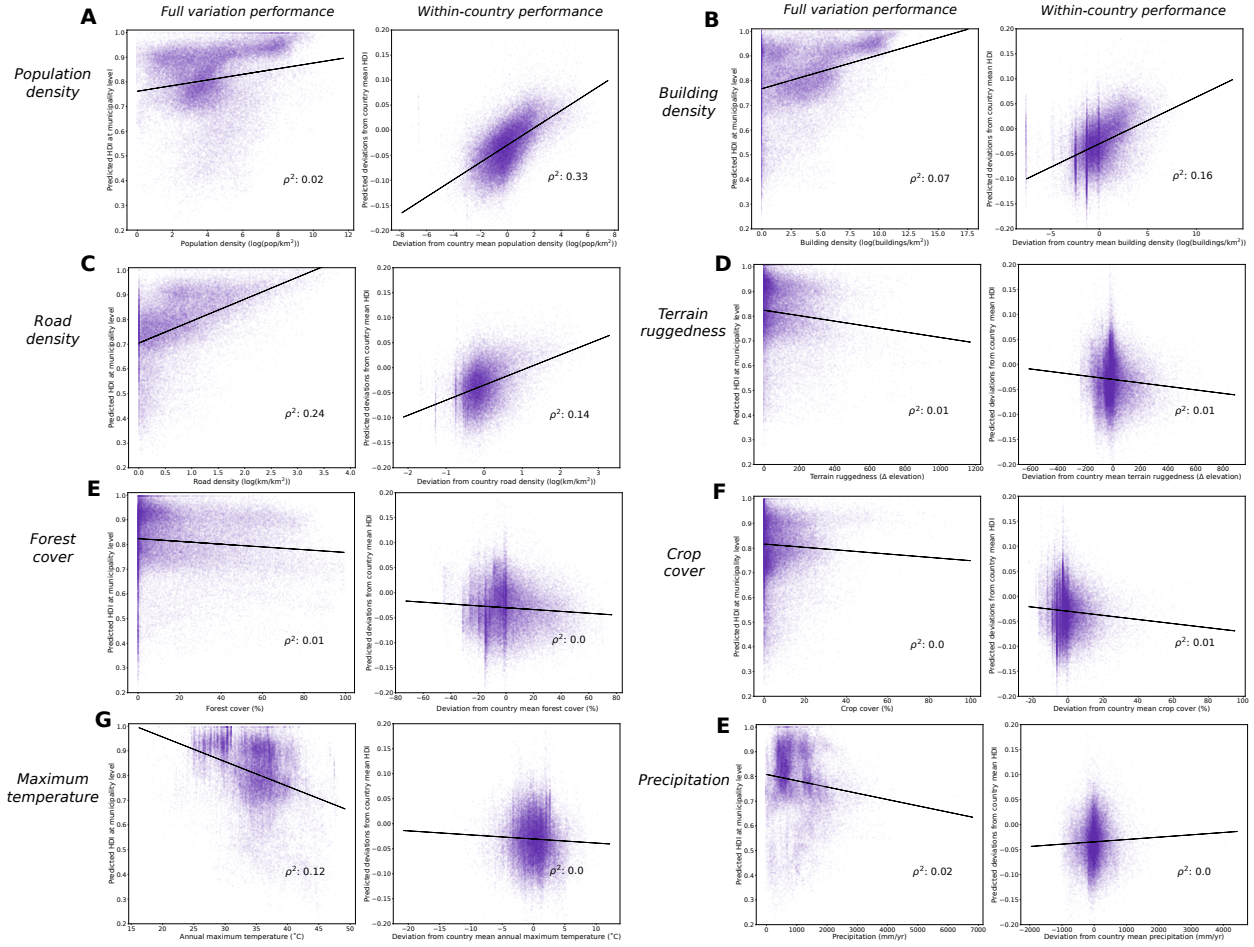

Figure S9: **Municipality-level HDI estimates can be partially predicted by other variables estimated at global scale.** Each point in each scatter represents the estimated HDI value and the value of another globally available variable for a municipality. See Supplementary Information S4 for details.

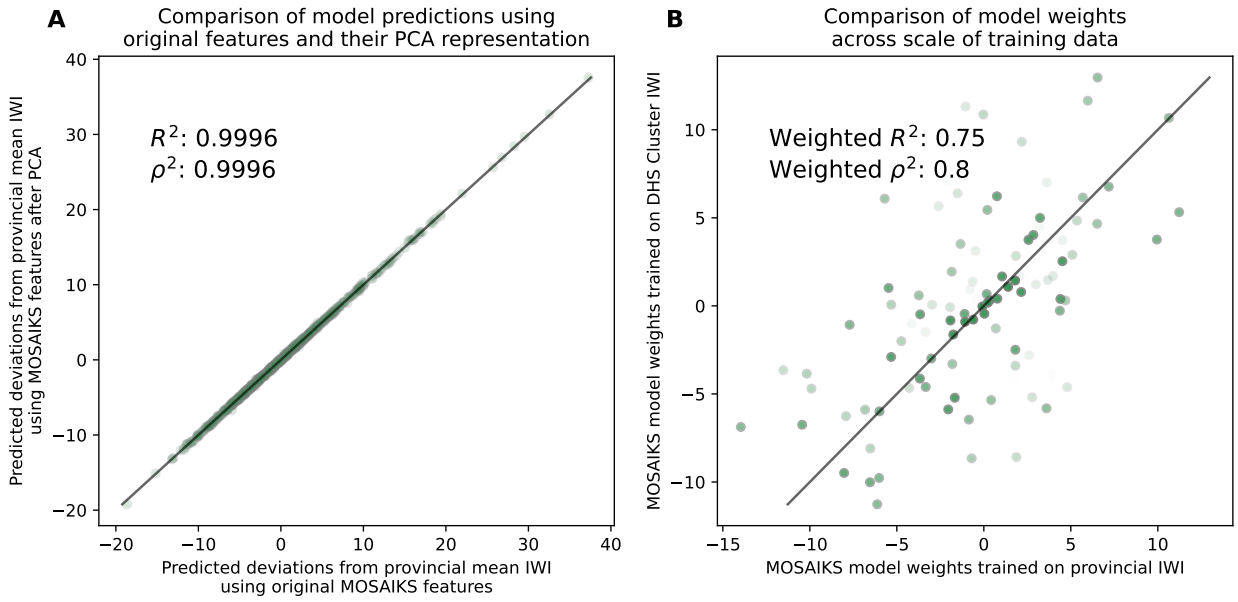

Figure S10: **Model weights for the IWI model are similar when trained at the provinicial or DHS cluster level. (A)** We show that estimates of IWI are nearly identical from the model with reduced dimensionality and the primary IWI model specification. **(B)** Each point represents the contribution to the model of a single feature. The points are colored by their rank order importance after principle component analysis. R^2 and ρ^2 are weighted by the explained variance under PCA. See Supplementary Information S5 for additional details.

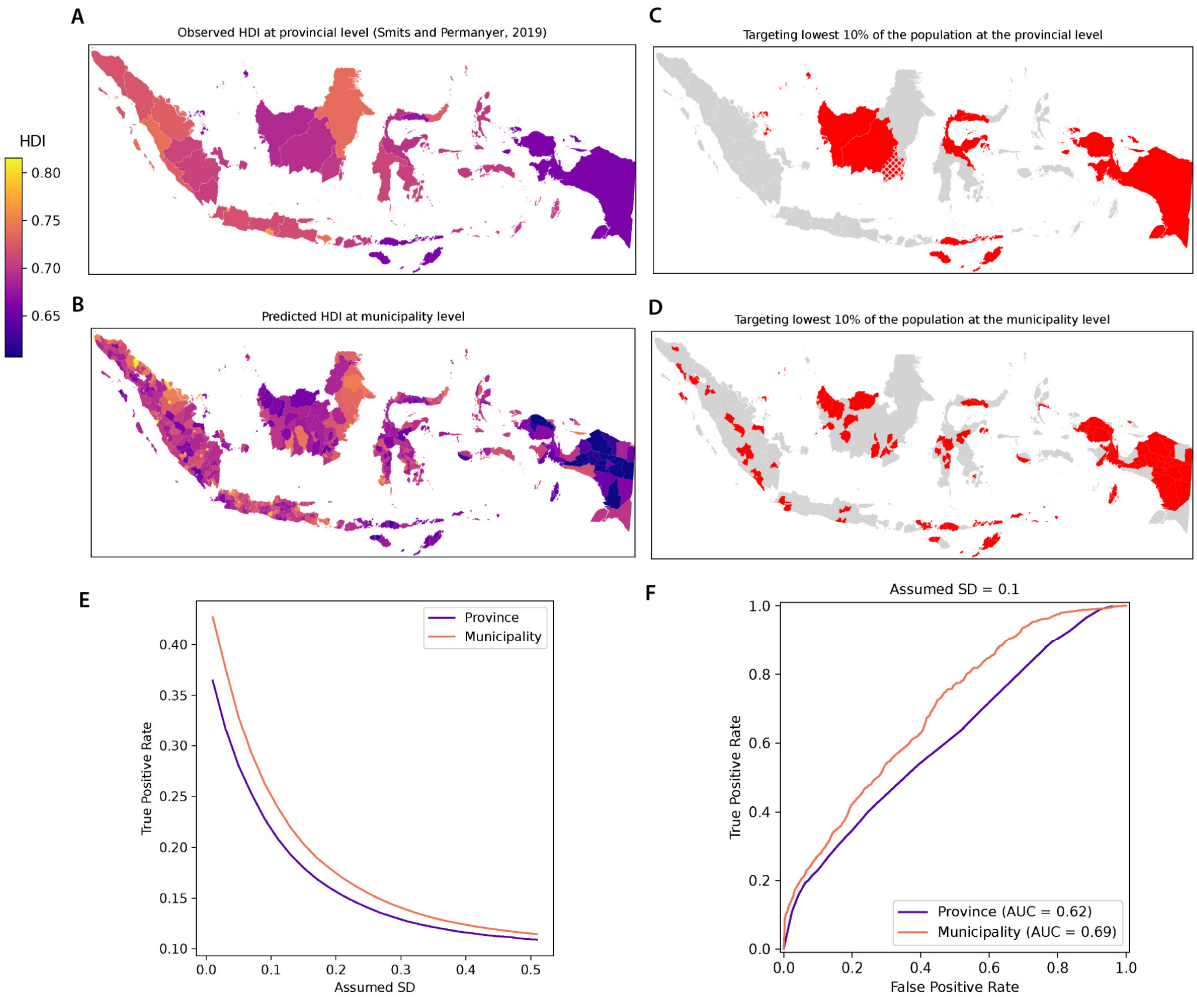

Figure S11: **Additional illustrative application: targeting policy in Indonesia.** Identical analysis as in Figure 5 using data from Indonesia. (**A**) HDI for 2019 at the province level of observation (7) (**B**) HDI estimates for 2019 at the municipality level produced in this paper. (**C**) Lowest HDI provinces that would be targeted until 10% of the country's population is reached. (**D**) Lowest HDI municipalities that would be targeted until 10% of the country's population is reached. Hashing in (**C**) and (**D**) shows the marginal province and municipality that would be partially targeted. (**E**) Targeting accuracy (true positive rate) as a function of the assumed standard deviation of HDI within each municipality. (**F**) ROC curves illustrate the degree of improvement that comes with targeting at the municipality level relative to the province level (assumed SD of 0.1 within each municipality).

S1 Supplementary discussion of the co-production of the analysis and estimates of HDI

This analysis and the associated estimates of the UN HDI were co-produced by a team of academics and practitioners aiming to develop measures of HDI that reflect the understanding and needs of both practitioners and policymakers (*68*, *69*). The effort was initiated by researchers and practitioners from the United Nations Development Programme (UNDP), coordinated by Dr. Heriberto Tapia, head of research for the Human Development Report Office and a co-author of this study. The Human Development Report Office is responsible for producing and publishing official HDI statistics for the United Nations. Members of the UNDP approached academic members of the team in 2021, with the goal of increasing the resolution of the HDI globally based on the growing demand for consistent disaggregated data following an impartial global standard. The combined team (practitioners and academics) jointly refined a project design and secured grants from third-parties and funding from the UNDP. Project implementation, writing, and dissemination was a joint and iterative process involving all team members.

Throughout the project, the team worked deliberately to ensure mutual knowledge transfer, with the academics learning what applications and aspects of modeling are most important to practitioners, and the practitioners learning the technical aspects of the model's development. The practitioners and academics collectively identified model performance as a primary priority, with model generalizability, simplicity, transparency and interpretability as additional design goals. Motivated by the results of this project, the UNDP is engaging in follow-on work that predicts the United Nations Multidimensional Poverty Index using the methods developed here. Academic members of our team are supporting and advising this follow-on work in an effort designed to solidify the transfer of this technology to UNDP and to build UNDP's technical capacity internally.

The UNDP team is particularly interested in satellite-based measurements of HDI for three reasons. First, satellite-based measurements adhere to relatively impartial technologically-led standards that facilitate international comparisons, a key attribute of UN indicators. Second, the integration of satellite imagery with machine learning creates new opportunities for using the HDI in policymaking. In the short term, this new data can facilitate planning and resource allocation within countries. In the medium term, these estimates should become a key input for analyzing the interaction between socioeconomic and geophysical data in studying the effects of climate change on people. This will be crucial for designing a new generation of policy responses through a global lens. Third, the availability of a database and a replicable method for generating these estimates is expected to address equity issues.

In a rapidly changing world, there is an increasing need for information for decision-making. Wealthier countries have high-quality surveys and the resources to power a data revolution. In contrast, poorer countries and communities, where data is most needed, face significant constraints in both surveys and analytical capacity. The availability of both a database and a standard, replicable method for generating these estimates is expected to be a valuable global public good.

S2 Supplementary discussion of model performance

Model performance across regions Analyzing the performance of MOSAIKS models across space, we find that performance tends to be the highest in low income regions, especially sub-Saharan Africa, where such measurements are likely to be of greatest value (Figure S7A,C). Specifically, we find that MOSAIKS models explain more variation of province-level HDI deviations from the country mean in areas of low human development (HDI < 0.6, $R^2 = 0.57$) than in areas of medium or high human development (HDI > 0.6, $R^2 = 0.43$, Figure S7A). We also see this pattern in predictions of DHS cluster-level IWI deviations from the province mean, with performance increasing monotonically from $R^2 = 0.09$ for countries with the highest HDI values to $R^2 = 0.52$ for countries with the lowest HDI values (Figure S7B,D). This improved performance may be due to increased variance of HDI and IWI values within these countries (Figure S7E-F), which provides more variation to exploit during model training. Alternatively, variations in wellbeing being may be relatively easier to see from satellite imagery in areas with lower human development. Relatedly, model performance for both HDI and IWI is higher in regions with higher inequality, highlighting the particular value of these estimates in regions of high inequality ($p < 0.01$ for both HDI and IWI; pearson's ρ is 0.33 for HDI and 0.53 for IWI comparing the within-country standard deviation of provincial values and the within-country predictive performance, measured by ρ^2).

Motivated by these differences in performance across regions, we test whether the relationship between image features and HDI varies spatially. We do so by training an additional model that allows the MOSAIKS features to have a continent-specific relationship with provincial HDI. We observe a small drop in within-country performance relative to the baseline model (from $R^2 = 0.52$ to 0.49), indicating that increasing the complexity of the model in this way does not improve out-of-sample performance.

Value from combining daytime and nighttime imagery The MOSAIKS-based approach can use image features from multiple sensors simultaneously when training models,

a property that is used throughout this analysis to predict HDI from both daytime and nighttime imagery. Analyzing the performance of MOSAIKS models based on the type of satellite imagery used, we find that daytime and nighttime imagery together explain 7% more variation in provincial HDI deviations from the country mean than does nighttime imagery alone, improving model fit by 16% from $R^2 = 0.45$ to 0.52 (Table S3). This improved performance from using daytime and nighttime imagery together is strongest in regions of low human development (HDI < 0.6) (Figure S7A), consistent with a previous finding that models using daytime imagery outperform models using nighttime imagery when predicting assets of the poorest populations in five African countries (*20*). Analyzing model performance for each component of HDI, we see that the improved performance predicting HDI using daytime and nighttime imagery stems from improved or comparable performance predicting each component of HDI (largest change in R^2 is 0.12 for expected years of schooling, smallest is no change in R^2 for life expectancy, Table S3).

In exploratory analysis of subsamples, we have observed that in several middle income countries with HDI between 0.7 and 0.8 (including those where we have municipal variation in HDI), within-country R-squared is sometimes higher when models are trained on nighttime imagery alone, relative to models that combine MOSAIKS and nighttime imagery. However, for other parts of the HDI distribution and on average, as described above, MOSAIKS features appear to add performance to the full model. We currently cannot explain why a nightlights-only model performs best for these particular countries when predicting HDI, and we do not know if this phenomena extends to non-HDI outcomes. In future work, we hope to present a more complete analysis that more fully evaluates how data from many satellite sensors, including others not used in this analysis, can be effectively and efficiently combined in SIML applications.

Model performance training at the country-level To evaluate performance in an extremely data-limited setting, we re-train our model using only country-level data. Despite a low number of training observations (N = 85 to 170 across experiments) these models maintain 36% to 54% of the performance of our preferred models trained using provincial deviations from the country mean (N = 862 to 2,852) when evaluated on the relative ordering of predicted and observed values using ρ^2 in all experiments (Table S1). This indicates that our approach can achieve competitive predictive performance detecting locations with relatively higher and lower HDI even when trained on few and coarse observations of the variable of interest. Performance predicting the exact level of HDI is lower, especially when evaluated within-country, likely due to the large difference in the magnitude of HDI variation across countries versus within countries, discussed above (Figure 2A,B).

S3 Supplementary analysis of temporal changes in the local distribution of HDI

We focus this manuscript on estimating fine-resolution spatial variation in HDI because existing provincial measures do not resolve within-province differences in human wellbeing, and because spatial variation in HDI is generally larger than temporal variation. We construct fine-resolution global estimates of HDI from 2012-2021 by combining time-varying estimates of provincial HDI from (*7*) with time-constant estimates of the local (i.e. within-province) distribution of HDI based on satellite imagery. Here, we conduct an additional experiment where we test whether allowing HDI to vary over time locally might improve model performance. Specifically, we compare our primary estimates (Figures 2 A-B and S8 A-B), which are made assuming a time-constant distribution of provincial HDI with another set of estimates that are made allowing for a time-varying distribution of provincial HDI (Figure S8 C-D). We find that these two approaches have near identical performance predicting historical HDI from 2012-2021. This motivates our use of the simpler model that assumes an approximately time-constant distribution of HDI as our primary specification when constructing municipal and grid-level estimates. Details of this analysis are described below.

To allow estimates of the local variation in HDI to change over time, we estimate a model similar to our primary specification (Equation S1a) but extended to train on and predict a global panel of HDI at the provincial level from 2012-2021. This model specifies HDI in each province and year as a linear function of VIIRS NL features for the relevant year and the MOSAIKS features from 2019. This allows the time-varying NL features in the model to predict changes over time in the distribution of provincial HDI. We do not allow the MOSAIKS features for the high-resolution visual imagery to change over time due to image data availability and computational limitations. In the models that allow for both time-constant and time-varying local HDI, we mean-anchor provincial estimates at the known national mean value for each year. Model evaluation using spatial cross-validation is conducted in the same way as in Figure 2, except that instead of evaluating on observations from a global cross-section of 2019 provincial data we evaluate on a global panel of provincial observations from 2012-2021.

We find that our primary model performance predicting a global cross-section of 2019 data (Figure 2 A-B) is similar to the performance of the same model predicting a global panel of HDI from 2012-2021 (Figure S8 C-D), as well as the performance of a similar model that allows for time-varying changes in local HDI predicting a global panel of HDI from 2012-2021 (Figure S8 A-B). Both the model assuming time-constant HDI estimates and the model

allowing for time-varying HDI estimates achieve an $R^2 = 0.96$ explaining provincial variation in the entire sample from 2012 – 2021. The former achieves an $R^2 = 0.53$ explaining within-country variation in the same sample, and the latter achieves an $R^2 = 0.54$ – an improvement of 2% from allowing the NL features to change over time. The model performance evaluated for each year is also similar for both approaches – with R^2s ranging between 0.96–0.97 for explaining variation across all provinces, and R^2s between 0.52–55 for explaining provincial variation within countries.

The similarity in performance between the model that uses time-constant and time-varying local HDI estimation suggests that the local-scale variation we predict with satellite imagery (here, within-country variation) is nearly constant over time for the period of observation (2012-2021). This time-consistency is further evidenced by the relatively stable performance predicting within-country variation each year, using models both with time-varying and time-constant NL features. Indeed, after removing country-specific national time trends, the spatial variation in HDI within each country across provinces is 60× larger than the respective temporal variation. The relatively large amount of variation in HDI across space and the relative time-consistency of the local pattern of HDI both motivate the focus of this analysis on increasing the spatial resolution of HDI globally.

S4 Supplementary discussion of model transparency and interpretability

The models employed in this analysis were designed to maximize predictive performance, based on co-development with our team members at the United Nations Development Programme (Supplementary Information S1) and feedback from other practitioners, who generally identified model performance as the most important property of the model. Other important aspects of model development are model transparency and interpretability (*57*) (Figures S1 and S9).

Model transparency Model transparency is the ability to clearly convey the model, its design, and its estimation (*57*). Transparency is a strength of the MOSAIKS approach used and developed here. The features are can be precisely described mathematically (*12*) and conveyed simply: each feature measures the similarity between a small patch of imagery and the image of interest, where similarity is measured using a moving dot product (i.e., convolution) of the patch over the imagery. This is illustrated in the cartoon developed in

Figure S1, which shows how three example patches capture the amount of trees, roads, and lakes in imagery. One explanation of how the MOSAIKS approach works is that the model predicts HDI in new areas where HDI is not known by assigning the (weighted) average value of HDI from places where HDI is known that look similar in the imagery (see Supplementary Note 2.3 in (*12*) for details). The linear model structure is also straightforward, especially compared with more complicated structures like a deep convolutional neural network or gradient boosted decision tree. We relate the image features to HDI using a linear ridge regression, which, unlike many more complex models, has an analytical solution. The simplicity and transparency of the MOSAIKS approach makes it straightforward to fully describe, understand and use (*12*). The work in this manuscript extends the ability of the MOSAIKS approach to learn the relationship between imagery and an outcome of interest using labels from any set of political boundaries, while maintaining a transparent design.

Model interpretability Model interpretability is the ability to understand what aspects of the input data are responsible for model predictions (*57*). To better interpret what aspects of the imagery are captured by our HDI estimates, we examine whether our municipality HDI estimates are correlated with known variables that can be constructed at the municipality level (Figure S9). Specifically, we construct municipality-level estimates for population density, road density, building density, terrain ruggedness, forest cover, crop cover, maximum temperature, and precipitation. Data for each of these variables is publicly available. We create a measure of each variable at the municipal level using the source data product and our municipality (ADM2) shapefile from geoBoundaries (*65*).

Population density Municipal population density data were constructed as the population count from GHS-POP divided by the area of the municipality (*62*).

Building density Building density data at the municipality level were constructed using the gridded building data product from Microsoft Bing Maps ($\approx 0.2 \times 0.2$ degree resolution) (*70*). This data is relatively coarse compared to the municipality size, so we assume a uniform distribution of buildings within the source raster when averaging building density over municipal polygons. We take this approach for all raster datasets used in this interpretability analysis.

Road density Municipal road density estimates were constructed by calculating the length of road in each municipal polygon and dividing by the polygon area. Road data are from Open Street Maps (*71*). To facilitate computation we calculate road density over a 1%

sample of global land area before averaging to the municipal level (≈1 km^2 sampled every ≈10km in the North-South and East-West directions). Municipality observations that do not contain a sampled road density observation are dropped from this analysis.

Terrain ruggedness Municipal ruggedness data were constructed using the global raster created by Shaver et al. (*72*) at 1km × 1km resolution. In this dataset "the ruggedness of any given 1 km^2 area is determined by measuring how the average elevation of that area differs from all those of neighboring 1 km^2 areas" (*72*). Municipal values are calculated as the average of ruggedness values within each municipality.

Forest cover Forest cover data come from the Hansen et al. data product (*73*), which is available at 30m × 30m resolution. Municipal estimates are calculated as the average forest cover over the municipal polygon.

Crop cover Crop cover data come from the Ramankutty et al. data product available at 0.08 × 0.08 degree resolution (*74*, *75*). For each municipality we calculate the fraction of the polygon area covered by cropland.

Maximum temperature We construct a measure of typical maximum annual temperature from the Climate Prediction Center gridded data product at 0.5 × 0.5 degree resolution (*76*). Specifically, we retrieve roughly a decade of daily data (2007-2018) and calculate the maximum temperature observed in each grid cell each year. We then take the average across years and over each municipal polygon.

Precipitation Precipitation data are from the NASA IMERG annual average precipitation data product at 0.1 × 0.1 degree resolution (*77*). We calculate municipal values as the average of gridded values over each municipal polygon.

To estimate the fraction of variance explained by each of these variables individually we execute a simple linear regression of our municipal HDI estimates on each variable. Within-country estimates are calculated after demeaning the HDI estimates and each variable by country. To estimate the fraction of variance explained by all of these variables together we execute a multiple linear regression of our municipal HDI estimates on all variables together. Before estimating this latter regression we used mean-imputation to fill in any missing values.

S5 Supplementary discussion assessing consistency of model weights across spatial scales

A key feature of our approach is that it can be trained on and make predictions for units of arbitrary shape and size. We employ this approach to train on global provincial HDI data and make predictions of HDI for global municipalities and a $0.1^\circ \times 0.1^\circ$ grid. For this approach to be effective, model weights estimated at the provincial level, must be able to make skillful predictions at the municipal and grid levels.

One way to understand how this approach works is to see that, in a linear model, the relationship between aggregated outcomes and aggregated features should be similar to the relationship between disaggregated outcomes and features. This is illustrated in Equations 2-3 of the main text. Rolf et al (*12*) provide additional mathematical explanation for relating predictions made using MOSAIKS features at image and sub-image scales. A primary goal of this manuscript is to propose that this approach can be used to address the challenge of limited training data in remote sensing applications by allowing for training on irregularly structured and sized observations – which is not discussed in ref. (*12*) – and to empirically test whether this works in practice. The primary evidence supporting this are the downscaling tests reported in Figure 2 and Table S1.

Here, we additionally explore the question of *why* the approach works by empirically testing whether model weights estimated at aggregated and disaggregated scales are similar. To do so, we compare model weights between models of IWI trained at the provincial (N = 862) and DHS cluster (N = 51,996) scales. We use IWI for this experiment because there are a large number of aggregated and unaggregated observations that span the same spatial extent.

A challenge in designing this experiment is that MOSAIKS features are correlated with each other, and there are a large number of features relative to the number of training observations. This means that the same information could load onto different features even when training and retraining at the same scale if we do not introduce additional constraints to the feature set. Put another way, different sets of model weights could give the same predictions and represent the same relationship between the imagery and outcome of interest. This is not an issue in our main application, since the set of weights obtained at an aggregated scale will remain valid if applied to a disaggregated scale, and vice versa. However, there is no guarantee that the same weights will be obtained if models are independently fit at both scales, since there are multiple valid ways to represent the data using the model features. Thus, for this experiment, we first transform our features into an orthogonal basis.

We use Principal Components Analysis (PCA) to project the MOSAIKS features into

a feature space with independent (i.e., uncorrelated) features that contain the same information as the original features, following ref. (*17*). PCA can also be used to reduce the dimensionality of the feature space, which can aid interpretation in this setting by focusing on features that explain most of the variation in the imagery, and thus likely in the outcome of interest. We find that 100 PCA features explain $> 99.9\%$ of the variation in the original 4,000 MOSAIKS features in this context, and that a model trained using these 100 PCA features provides essentially identical predictions to a model trained using the 4000 original features (Figure S10A). This PCA model is thus practically identical to our MOSAIKS-based model but with independent features that represent an orthogonal basis. We use these orthogonalized and rotated MOSAIKS features to analyze whether model weights are consistent when training at different levels of aggregation.

Using these orthogonal MOSAIKS features, we find that model weights estimated using the (aggregated) provincial data are very similar to model weights that are independently estimated using (disaggregated) DHS cluster data (R^2=0.75, Figure S10A). Note that when calculating R^2 in this setting we weight by the fraction of variance in the MOSAIKS features each component explains, so that greater weight is placed on features that explain more variation in the MOSAIKS features, and likewise, in IWI. The high correspondence between weights estimated at aggregated and disaggregated scales indicates that the same satellite information is being used in the same way to predict IWI at both scales. This helps to explain how our approach is able to achieve skill in the downscaling applications illustrated in Figure 2 and Table S1.

S6 Supplementary methods

Note that in the supplementary methods we use the subscript p to refer to provincial or first-level administrative regions; and the subscript m to refer to municipality or second-level administrative regions. We use the subscript c to denote observations at the country level.

S6.1 HDI model training

Within-country model training Because our focus is explaining subnational variation in HDI, we specifically train our primary model to predict within-country deviations of HDI. To do this, we first demean subnational observations by country and then train a model to use imagery to predict these residualized deviations. Specifically, we transform observed ADM1 HDI for province p (HDI_p^{ADM1}) into the deviation of this value from the country

mean HDI ($\widetilde{HDI}_p^{ADM1}$). We then solve a ridge regression to predict $\widetilde{HDI}_p^{ADM1}$ based only on provincial daytime ($\widetilde{\mathrm{X}}_{MOSAIKS,p}^{ADM1}$) and nightlight ($\widetilde{\mathrm{X}}_{NL,p}^{ADM1}$) features that have been similarly residualized relative to the country mean values for these variables. We learn the model

$$\widetilde{HDI}_p^{ADM1} = \beta_0 + \beta_1 \, \widetilde{\mathrm{X}}_{MOSAIKS,p}^{ADM1} + \beta_2 \, \widetilde{\mathrm{X}}_{NL,p}^{ADM1} + \epsilon_p \tag{S1a}$$

where :

$$\widetilde{HDI}_p^{ADM1} = HDI_p^{ADM1} - \sum_{p \in c} \frac{HDI_p^{ADM1}}{N_c} \tag{S1b}$$

$$\widetilde{\mathrm{X}}_{MOSAIKS,p}^{ADM1} = \mathrm{X}_{MOSAIKS,p}^{ADM1} - \sum_{p \in c} \frac{\mathrm{X}_{MOSAIKS,p}^{ADM1}}{N_c} \tag{S1c}$$

$$\widetilde{\mathrm{X}}_{NL,p}^{ADM1} = \mathrm{X}_{NL,p}^{ADM1} - \sum_{p \in c} \frac{\mathrm{X}_{NL,p}^{ADM1}}{N_c}. \tag{S1d}$$

Here, N_c is the number of provinces in country c. Note that we restrict predictions from this demeaned model to be between the observed minimum and maximum HDI deviations from the country mean.

Anchoring to country means via re-centering To evaluate full variation performance using the within-country model (Table S1, col. 1-2) we need HDI predictions in "levels" rather than predicted deviations from the country mean. To construct predicted HDI values in "levels" we anchor our estimates to country means, since they are observed and used in the estimation procedure. Practically, this means we add the country mean HDI, which was subtracted from the observations before model training, back onto the predicted deviations:

$$\widehat{HDI_p^{ADM1}} = \widehat{\widetilde{HDI}_p^{ADM1}} + \sum_{p \in c} \frac{HDI_p^{ADM1}}{N_c} \tag{S2}$$

Note that it is not necessary to implement this procedure when evaluating within-country performance.

Province and country model training In Table S1, we additionally report performance for models trained on province and country-level data directly. Unlike the within-country model, these models are trained on values in "levels" instead of deviations from the country mean. In these experiments, we learn the models:
Province model:

$$HDI_p^{ADM1} = \beta_0 + \beta_1 \mathrm{X}_{MOSAIKS,p}^{ADM1} + \beta_2 \mathrm{X}_{NL,p}^{ADM1} + \epsilon_p \tag{S3}$$

Country model:

$$HDI_c^{ADM0} = \beta_0 + \beta_1 \ \mathrm{X}_{MOSAIKS,c}^{ADM0} + \beta_2 \ \mathrm{X}_{NL,c}^{ADM0} + \epsilon_c \tag{S4}$$

We do not apply a mean-anchoring procedure with these models as their predictions are already in "levels" rather than predicted deviations. Note that 20 of the 179 total countries do not have subnational data (e.g., Qatar) and that these 20 country-only observations are included in both province and country models.

S6.2 Downscaling validation with IWI

Labels IWI is similar to the wealth index reported in DHS surveys, except that it was created to be comparable across countries (*36*). IWI data are available both for provincial polygons, which we use for training, and for DHS clusters, which we use for evaluation. For each survey cluster, DHS provides coordinate points associated with the cluster centroid. To protect privacy, the actual GPS coordinates of the center of each cluster are randomly displaced by up to 2km for urban clusters and up to 5km for rural clusters, with a random 1% of rural cluster coordinates displaced by up to 10km. According to DHS, the displaced coordinate is guaranteed to fall within the same DHS-provided administrative boundaries as the true cluster centroid. To map these point observations to administrative polygons, we spatially buffer urban cluster coordinates using a 2km radius and rural cluster coordinates using a 10km radius. We then clip these buffers to the finest DHS-provided administrative boundaries that are available.

Training We train within-country, province level, and country level IWI models following the structure of models for HDI (Methods Section S6.1). Provincial IWI observations are denoted IWI_p^{ADM1}.

The within-country IWI model, our preferred model specification, takes the same form

as Equation S1a:

$$\widetilde{IWI}_p^{ADM1} = \beta_0 + \boldsymbol{\beta}_1 \, \widetilde{\mathrm{X}}_{MOSAIKS,p}^{ADM1} + \boldsymbol{\beta}_2 \, \widetilde{\mathrm{X}}_{NL,p}^{ADM1} + \epsilon_p \tag{S5a}$$

where :

$$\widetilde{IWI}_p^{ADM1} = IWI_p^{ADM1} - \sum_{p \in c} \frac{IWI_p^{ADM1}}{N_c} \tag{S5b}$$

Note that $\widetilde{\mathrm{X}}_{MOSAIKS,i}^{ADM1}$ and $\widetilde{\mathrm{X}}_{NL,i}^{ADM1}$ are the same feature matrices defined in Equation S1c and S1d but with a different number of observations due to differing availability of outcome data.

Prediction We evaluate the IWI model performance at a finer resolution than it was trained. We use the trained provincial model (Equation S5a) to produce predictions of IWI at the DHS cluster level and compare those predictions to the cluster-level IWI measurements from the GDL, which were not used for model training. We calculate DHS cluster-level features in the same way as for the other administrative polygons.

To make predictions of IWI deviations from the country mean at the DHS cluster level using the within-country model trained on provincial deviations from the country mean, we multiply model weights with the demeaned DHS cluster-level satellite features:

$$\widetilde{\widehat{IWI}}_d^{DHS} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1 \, \widetilde{\mathrm{X}}_{MOSAIKS,d}^{DHS} + \hat{\boldsymbol{\beta}}_2 \, \widetilde{\mathrm{X}}_{NL,d}^{DHS} \tag{S6}$$

where d indexes DHS cluster and $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}_1$, and $\hat{\boldsymbol{\beta}}_2$ are estimated in Equation S5a. Tildes denote that these predictions are predicted deviations from the country mean. In our within-country IWI model, we demean DHS cluster-level satellite image features by the same country average feature values as in the training procedure:

$$\widetilde{\mathrm{X}}_{MOSAIKS,d}^{DHS} = \mathrm{X}_{MOSAIKS,d}^{DHS} - \sum_{p \in c} \frac{\mathrm{X}_{MOSAIKS,p}^{ADM1}}{N_c} \tag{S7a}$$

$$\widetilde{\mathrm{X}}_{NL,d}^{DHS} = \mathrm{X}_{NL,d}^{DHS} - \sum_{p \in c} \frac{\mathrm{X}_{NL,p}^{ADM1}}{N_c} \tag{S7b}$$

where we note that these averages are constructed by averaging province-level features, but

have similar values that averages of nationally-representative sets of cluster-level features would have.

Anchoring to provincial means via re-centering To construct estimates of cluster-level IWI in levels ($\widehat{IWI_d^{DHS}}$), we anchor predicted cluster-level deviations from the country mean ($\widehat{\widetilde{IWI_d^{DHS}}}$) to the known provincial value (IWI_p^{ADM1}) using a provincial level adjustment:

$$\widehat{IWI_d^{DHS}} = \widehat{\widetilde{IWI_d^{DHS}}} + \underbrace{IWI_p^{ADM1} - \sum_{d \in p} \frac{\widehat{\widetilde{IWI_d^{DHS}}}}{N_p}}_{\text{centers DHS clusters to known provincial values}} \tag{S8}$$

Here, N_p denotes the number of DHS clusters contained by ADM1 polygon p, and IWI_p^{ADM1} denotes the observed ADM1-level value for polygon p. This anchors the mean of our DHS cluster-level predictions within each provincial polygon to the respective known province value used in training.

S6.3 Downscaling validation using nighttime lights as labels

In our analysis of the downscaling perfomance of our approach, we design an experiment in which NL are used as *labels* and are *not used as features* (Figure S2). This experiment is useful because it is the only validation experiment where the ground truth data are available globally and at municipal resolution. Thus, this experiment allows us to evaluate predictions at a downscaled resolution for the entire globe using a procedure that mirrors how we will generate downscaled HDI estimates (such global high resolution labels do not exist for our other outcomes). We do not expect NL predictions to be perfect proxies for HDI data in this regard, but if NL can be downscaled successfully, it provides support for the *procedure* we use to downscale HDI.

Labels We use population estimates from GHS-POP and fine resolution NL data from VIIRS to create a population-weighted average NL raidance at the province level. NL observations are population-weighted to mirror the construction of HDI, which is also population-weighted. Combining NL observations with population is also common practice when using NL as a development indicator (*41*, *43*, *44*). We construct municipality-level NL observations using a municipal (ADM2) shapefile from geoBoundaries (*65*), which links municipalities to provincial "parent" polygons.

We exclude Ireland from the geoBoundaries ADM2 dataset because Irish municipalities (ADM2 units) are so small that they alone represent 45% of the global municipality observations. Thus, they would be over-represented in global performance metrics relative to their size if not removed.

The vertical streaking patterns in the scatter plots in Figure 2H-J are caused by other countries that also have very spatially dense municipalities, though not to the same degree as Ireland. Because many within-province predictions are clipped at the observed minimum or maximum within-country deviation, this creates vertical streaking at the extremes in 2J. When the country-level mean values are added back, this results in vertical streaking at an arbitrary point along the x-axis in 2H-I. The three countries that mostly account for this effect are Great Britain ($\approx$ 9,000 units), Spain ($\approx$ 8,000 units), and Brazil ($\approx$ 5,000 units).

Training We train a model using only MOSAIKS features constructed from daytime imagery to predict NL:

$$NL = \beta_0 + \boldsymbol{\beta}_1 \, \mathbf{X}_{MOSAIKS} + \epsilon \tag{S9}$$

This model structure is broadly the same training procedure described in Methods Section S6.1 and in Equation 4; however, we do not include NL features when predicting average NL luminosity. NL is also now a vector of scalar NL observations rather than a matrix of features.

Prediction To generate municipal predictions, indexed by m, from the within-country model, we first create municipal predictions of NL deviations from the country mean. We demean $\mathrm{X}^{ADM2}_{MOSAIKS,m}$ by country by subtracting the country mean feature values and then multiplying the resulting demeaned features by the estimated model weights. This corresponds to what is done when evaluating downscaled IWI performance in Section S6.2 and shown in Equation S7a.

Anchoring to country means via re-centering When converting the predicted municipal NL deviations from the country mean ($\widetilde{\widehat{NL}}^{ADM2}_m$) into predicted municipal NL values in levels ($\widehat{NL}^{ADM2}_m$), we anchor values to the known country mean:

$$\widehat{NL}^{ADM2}_m = \widetilde{\widehat{NL}}^{ADM2}_m + \sum_{p \in c} \frac{NL^{ADM1}_p}{N_c} \tag{S10}$$

Note that we anchor fine resolution NL predictions to the known country mean rather than the provincial mean (following Equation S2 rather than Equation S8) because we find that

this substantially improves full variation performance. Most of the variation in nightlight luminosity occurs within countries, rather than between countries, which is considerably different from what we observe for HDI and IWI. Importantly, the choice to use a different re-centering procedure for NL does not impact the downscaled within-province performance (Figure 2J), which we believe provides the most important evaluation of downscaling performance.