Local GDP Estimates Around the World*

Esteban Rossi-Hansberg
University of Chicago

Jialing Zhang
Princeton University

April 28, 2026

Abstract

We use high-resolution spatial data to build a novel global annual gridded GDP dataset at 1°, 0.5°, and 0.25° resolutions from 2012 onwards. Our random forest model trained on local and national GDP achieves an R^2 above 0.94 for GDP levels and above 0.73 for annual changes in regions left out of the training sample. By incorporating diverse indicators beyond population and nighttime lights, our estimates offer more precise subnational GDP measurements for analyzing economic shocks, local policies, and regional disparities. We evaluate the precision of using a case study of COVID-19's impact on local GDP in China.

Analyzing the spatial structure of the world economy and its evolution requires measures of economic activity at high geographic resolutions. Unfortunately, obtaining worldwide detailed production data over time is particularly challenging because middle- and low-income countries lack reliable subnational statistics. The G-Econ dataset by Nordhaus (2006) provides essential spatial GDP estimates, primarily using population as a key predictor, but is limited to a few years (1990, 1995, 2000, 2005). Researchers seek datasets that span more recent and frequent years and move beyond population-based proxies. The work by Henderson, Storeygard and Weil (2012) was pivotal in demonstrating that nighttime lights (NTL) can serve as a reliable proxy for economic activity. The use of NTL data for GDP estimation in data-scarce regions involves various methods, including running regressions with coarser GDP data to generate finer-scale predictions (Vogel et al., 2024), directly using NTL as the outcome variable in regressions (Storeygard, 2016; Henderson, Squires and Weil, 2018), and employing light intensity as a weighting factor to distribute national GDP (Chen et al.,

*Esteban Rossi-Hansberg: earossih@uchicago.edu. Jialing Zhang: zhang.jialing@princeton.edu. We thank Reigner Kane, Sreyas Mahadevan, Jordan Rosenthal-Kay, and Julian Tsang for excellent research assistance and contributions during the initial stages of this project.

2022). It is well known, however, that nighttime lights data have limitations: saturation can lead to underestimation of urban production (Henderson, Storeygard and Weil, 2012), inconsistencies between satellite sensors over time pose challenges for comparing data across years, and sectors such as agriculture and forestry are often underestimated by NTL (Keola, Andersson and Hall, 2015).

There is an increasing demand for more accurate local GDP datasets to enhance the quantification of spatial models, as well as for reduced form analysis. To meet this need, studies have used machine learning techniques, combined with high-resolution datasets (Khachiyan et al., 2022). However, these machine learning approaches are typically restricted to regions with reliable subnational GDP data for training, and thus have limited applicability for middle- and low-income countries.

Building on this approach, we go a step further to develop a statistical prediction model that generalizes across heterogeneous regions and still accurately predicts local GDP for out-of-sample regions. Our model can provide reliable local GDP estimates that cover more years and also capture economic shifts and complex patterns, as we describe below.

To achieve this goal, we introduce several novel elements. First, alongside gridded population data (Bright et al., 2012–2021) and NASA's new unsaturated nighttime lights dataset (VIIRS VNP46A4 Black Marble product) (Román et al., 2018), we also integrate high-resolution land use data (Friedl and Sulla-Menashe, 2022), emissions data (Emissions Database for Global Atmospheric Research , EDGAR), and net primary productivity (NPP) data (Running and Zhao, 2021).[1] The unsaturated nighttime lights dataset addresses urban GDP underestimation issues and mitigates the challenges of data comparison across years as its sensors remain consistent over time. Integrating land use and NPP data improves GDP estimates in the agricultural sector. Emissions data capture more nuanced changes in GDP in response to economic shocks. This diverse set of variables enables more accurate out-of-sample GDP predictions, improves year-over-year GDP growth rate estimations, and captures economic shifts that single proxies may miss.[2]

Second, we use a random forest algorithm to build our statistical prediction model. This methodology excels at preventing overfitting by aggregating predictions from numerous weak learners (i.e., decision trees) while still capturing complex data patterns. We use it to estimate the function that maps predictor shares to GDP shares using subnational

[1]Net Primary Productivity (NPP) measures the amount of carbon captured by plants through photosynthesis minus the carbon they release via respiration. It serves as an indicator for measuring vegetation.

[2]Variable importance analysis confirms this: nighttime lights in cropland areas contribute primarily at the 1-degree level. CO_2 emissions from heavy industry are among the most important non-population predictors, particularly at the 0.5-degree and 0.25-degree resolutions. Land use data such as cropland and urban classifications interact with population to amplify its predictive power. For example, population in urban areas is far more informative than total population alone.

GDP data from North America, South America, Europe, Africa, and Asia. By focusing on GDP share instead of absolute GDP levels, the model disregards country-level fixed effects and can generalize more effectively across regions. To avoid overly optimistic out-of-sample predictions, we use Group k-Fold Cross-Validation. This method keeps data points from the same country together, preventing them from being split into training and testing sets. This ensures that when we select the best model, we do not choose the one that appears to perform well due to shared data dependencies between training and testing or because GDP share values do not vary much.

Our model achieves out-of-sample R^2 values above 0.94 for GDP levels and 0.73 for annual changes. We further validate the model by applying it to China, another country excluded from the training sample, to assess its accuracy in the case of an unprecedented shock such as COVID-19. The model achieves R^2 values above 0.94 for GDP levels and above 0.52 for annual changes across pre-COVID, COVID, and post-COVID periods. Furthermore, unlike the G-Econ dataset, where population is the primary GDP predictor, our model identifies a convex relationship between population and GDP at all three spatial resolutions, consistent with the presence of agglomeration economies. Additional analyses presented in the Online Appendix demonstrate the model's strong in-sample fit, consistency across resolutions, predictive accuracy for future years, and robustness under various checks.

The remainder of this paper is structured as follows. Section 1 describes the primary data sources, data construction processes, and model training procedures. Section 2 presents the results and evaluates the model performance. Section 3 explores potential applications of our datasets and concludes. All data are accessible for download at `http://bfidatastudio.org/gdp`.

1 Data and Methods

1.1 Data

We use three primary types of data: remote sensing data, model-processed datasets from other sources, and GDP datasets from international organizations and official agencies. See Online Appendix Section 1 for a detailed description of the data sources.

The remote sensing data include nighttime lights (NTL), land use, and net primary productivity (NPP) data. The nighttime light data comes from NASA's VIIRS VNP46A4 Black Marble product. It provides uncensored nighttime light intensity at 15 arc-seconds (approximately 450 to 500 meters) with consistent satellites across years and removes noise and temporary lighting. We further process the data by 1) excluding the lights within a

0.2° cell radius centered around known gas flaring locations with positive gas flaring volumes sourced from the Global Gas Flaring Data dataset in the Global Flaring and Methane Reduction Partnership (GFMR) community (Global Gas Flaring Data, 2012-2023), and 2) excluding ocean and large inland water bodies before extracting the lights emissions in each cell.[3] Different land activities emit varying levels of light and may relate differently to GDP, so we also use NASA MODIS MCD12Q1 version 6.1 land cover dataset to separate lights by urban, cropland, and other land types. This dataset has a spatial resolution of 500 meters. We also include land cover data as separate predictors in our statistical model. To address the limitations of nighttime lights in capturing the agricultural sector, we supplement cropland land-use data with net primary productivity (NPP) data from the NASA MODIS MOD17A3HGF version 6.1 product. NPP measures vegetation with a spatial resolution of 500 meters.

Gridded population and CO_2 emissions data are model-processed datasets from other sources. Population data, derived from the LandScan Global dataset, spatially distributes national population counts at a 30 arc-second resolution (approximately 1 km) by applying country-specific machine learning models based on satellite imagery. The CO_2 emissions dataset comes from the Emissions Database for Global Atmospheric Research (EDGARv8.0), specifically the Global Greenhouse Gas Emissions dataset. It has a spatial resolution of 0.1 degrees. We categorize and aggregate sector-specific emissions into six groups: manufacturing combustion, heavy industry, and transportation, each further divided to separately account for fossil CO_2 sources and biofuel CO_2 sources. Alongside the above data, we include the terrain ruggedness index from Nunn and Puga (2012).

National GDP and national population data are sourced from the IMF World Economic Outlook Database, with supplementary data from the World Bank and UNdata for specific countries. Regional GDP data are mainly obtained from the OECD Regional Statistics dataset, complemented by additional data for some developing countries from Wenz et al. (2023) and official statistical agencies in Russia, Brazil, China, India, Kazakhstan, USA, Philippines, and Kyrgyzstan. For China, city-level data from the seven provinces with the highest GDP are collected from provincial statistical yearbooks. These city-level data are used exclusively to evaluate the model's ability to detect COVID-19 shocks and are not included in any model training.[4] Further details on data sources, processes, and countries included in the training sample can be found in Online Appendix Section 1.

[3] 3,067 largest lakes (area greater than or equal to 50 km^2) and 654 largest reservoirs (storage capacity greater and equal to 0.5 km^3) worldwide are excluded from the geometry shapefiles.

[4] China's GDP data is excluded from the training set primarily due to the challenges in data collection, including restricted overseas access to government websites, the need for manual compilation, and discrepancies between the China City Yearbook and individual province statistical yearbooks.

1.2 Random Forest Model Training

Training datasets are constructed at the cell level for each resolution (1°, 0.5°, and 0.25° using the EPSG:4326/WGS84 coordinate reference system), covering the years 2012 to 2022.[5] A separate random forest model is developed for each resolution. The random forest algorithm is selected for its ability to capture complex relationships and also mitigate overfitting by averaging the predictions of multiple decision trees (Breiman, 2001).

Predictor aggregation and conversion to shares. The predictor set includes nighttime lights (NTL), population, CO_2 emissions, net primary productivity (NPP), land cover shares, mean terrain ruggedness, and national GDP per capita. Each predictor is constructed by first aggregating raw pixel values to the cell level, then converting to shares relative to the parent area.

Because the raw predictor data are available at a pixel resolution finer than our target cell level, we aggregate the values of pixels k intersecting with cell i to compute the cell's absolute predictor value $\tilde{X}_i$. This is done by multiplying each pixel's density x_k by the intersected area A_{ki},[6]

$$\tilde{X}_i = \sum_{k \in \{k | k \cap i \neq \emptyset\}} x_k \times A_{ki}. \tag{1}$$

The units of x_k vary depending on the predictor. For nighttime light (NTL) data, x_k is measured in nanowatts per square centimeter per steradian ($n\text{Watts} \cdot \text{cm}^{-2} \cdot \text{sr}^{-1}$). For NPP, x_k is expressed in kilograms of carbon per square meter (kgC/m^2). For CO_2 emissions, x_k is in tonnes per square meter (t/m^2). For population, x_k represents the population count per square meter. For land-use data, x_k is an indicator variable that identifies a specific land-use type within the pixel.[7]

[5] The output resolution is chosen based on a trade-off among predictor availability, training data granularity, computational cost, and usefulness in economic applications. Among our predictors, the coarsest resolution is 0.1° (EDGAR CO_2 emissions data). The subnational income data used to construct cell-level GDP (i.e., county-level or equivalent administrative data) vary in spatial granularity: for developed countries, average subnational areas range from 695 km² (Belgium) to 29,668 km² (Norway), with a mean of approximately 6,779 km²; for developing countries, the range is 2,353 km² (Albania) to 70,691 km² (Mozambique), with a mean of approximately 25,143 km². See Online Appendix Section 1.2 Table 2 for a list of countries in the training sample and their average area of regional units used to construct cell-level GDP data for model training. Considering Earth's curvature, 1°, 0.5°, 0.25°, and 0.1° grid cells correspond on average to areas of roughly 12,300 km², 3,100 km², 770 km², and 123 km², respectively. Balancing these considerations, we produce estimates at resolutions of 1°, 0.5°, and 0.25°. The available subnational training data is not detailed enough to produce estimates at finer spatial scales.

[6] The areas are calculated based on a spherical approximation of Earth.

[7] For cells located along country borders, the cells are divided into segments based on the boundaries. The aggregated value for each segment is calculated separately to account for the portion of the cell that falls within each country.

After obtaining cell absolute values, we construct predictors as shares relative to country-level or state-level wherever available (Australia, Brazil, Canada, China, India, Kazakhstan, Mexico, Russia, USA). Expressing predictors as shares rather than absolute levels removes country-specific level effects, enabling the prediction model to generalize to countries not in our training sample. Using shares also naturally normalizes variables measured in different units to a common scale. Alongside the above shares, we include lagged shares to help predict growth rates.[8] In addition, the terrain ruggedness index is calculated as a cell mean, and national GDP per capita is also added to the model.

Construction of cell-level GDP shares. The model is trained using available county-level GDP data for developed countries and data available at a slightly broader administrative level for developing ones. These training countries span North America, South America, Europe, Africa, and Asia. Assuming uniform GDP per capita within counties or comparable areas, we calculate cell-level GDP as $y_i = \sum_{r \in \{r|r \cap i \neq \emptyset\}} \frac{y_r}{p_r} \times p_{ir}$, where $\frac{y_r}{p_r}$ denote GDP per capita of county r, and p_{ir} denotes the population at the intersection of county r and cell i.[9] As when calculating predictors' share, we calculate cell GDP share by country, and for countries with state-level GDP available (Australia, Brazil, Canada, China, India, Kazakhstan, Mexico, Russia, USA), we calculate state-level shares.

Cross-validation and hyperparameter tuning. In order to ensure that our random forest prediction model generalizes well to unseen countries and does not overfit to the training data, we use cross-validation to tune three hyperparameters: terminal node size, number of candidate variables at each split, and number of trees. To predict GDP shares for cells in entirely out-of-sample countries, we cannot use the traditional cross-validation approach of creating folds through splitting data points, as this could result in each fold containing the same cell across different years. If a cell's GDP shares remain stable, this approach may yield overly optimistic results on test folds while still failing for unseen cells in new countries. Instead, we randomly divide the training countries into five folds, using cells from four folds for training and cells from the remaining fold for testing to measure "out-of-training-sample" R^2 in log levels and in year-over-year log differences, for both developed and developing countries' cells (see Section 2.3 for results). Because the partitioning is done

[8] Since the Black Marble nighttime lights data constrain our data to start in 2012, no lag variable is available for that year; instead, we use the 2012 data as its own lagged value. Dropping the year 2012 data from our training sample has minimal impact on GDP predictions. See Online Appendix Section 12.3 for the comparisons.

[9] Again, for cells located along country borders or state borders (in cases where state-level GDP is used), the cells are divided into segments based on the boundaries and the aggregated value for each segment is calculated separately.

at the country level, no cell in the held-out countries appears in the training set, which eliminates spatial leakage from neighboring cells. Furthermore, all observations for a given cell across the full sample period (2012–2022) are assigned to the same fold, so the concern that performance may be inflated by the same cell appearing in both training and testing sets across different years does not apply. For each hyperparameter combination, this five-fold approach provides five "out-of-training-sample" R^2 values. We then calculate the mean R^2 across the five folds for each group (developed and developing) by taking a weighted mean of these two values based on the proportions of cells belonging to each group in the world. The hyperparameters are chosen to maximize the weighted R^2 of year-over-year log differences.[10]

Training of the final random forest model. The optimal hyperparameters selected via cross-validation are then used to train the random forest on the entire training dataset. This final random forest is used for all predictions and subsequent analyses. To address the imbalance in the training data between cells from developed and developing countries, we train the final random forest with weights that rescale the shares of these cells to reflect their actual proportions in the real world. The weights affect the probability of each cell being selected in the bootstrap sample to build the decision tree and allow a balanced representation of both developing and developed countries during training.[11]

1.3 Predictions and Post-adjustments

After building the random forest model, we obtain GDP predictions for all cells. For training cells, we use out-of-bag predictions to provide unbiased estimates. The out-of-bag estimate is calculated using predictions from trees that excluded the cell from their training. For out-of-training-sample cells, predictions are generated directly by the model. To avoid overestimating GDP per capita in sparsely populated cells, we censor the predicted cell GDP shares to zero for those with zero population density.[12] Then, GDP shares are rescaled

[10]Random forest models typically pick the hyperparameters that generate the lowest mean squared error (MSE) rather than focusing on year-over-year log differences. However, tuning based on year-over-year log differences, while resulting in a slight increase in MSE, can significantly improve the accuracy of log difference predictions. This adjustment has minimal impact on the accuracy of GDP level predictions. Refer to the Online Appendix Section 7.1 for results from models tuned using MSE and their comparison with the benchmark models presented in the paper.

[11]Refer to the Online Appendix Section 7.2 for a comparison showing that weights have minimal impact on model performance.

[12]Cells located along country or state borders may form irregular polygons instead of standard grid cells. For these polygons, censorship is applied at the polygon level. Thus, after aggregation into standard cells, some cells may still display positive GDP values. We also provide datasets where censorship is applied to cells with population densities below thresholds of 0.01, 0.02, or 0.05 individuals per km^2 of cell land area. All results in Section 2 use the dataset adjusted for cells with zero population density. See Online Appendix Section 4 for a count of cells affected by the adjustments.

within parent areas to sum to 1, and GDP predictions are obtained by multiplying these rescaled shares by the GDP of the parent area. Our procedure also allows us to obtain the standard deviation of predictions for each cell.[13]

2 Local GDP Predictions

2.1 Global GDP Distribution and Growth Patterns

This section presents the results of the 1-degree model trained using data from 2012 to 2022 for all available countries (excluding China) under the optimal hyperparameters, and then used to generate predictions for global GDP at a 1-degree resolution. Figure 1 presents GDP and GDP per capita results in both levels and changes. Grey cells are those with zero population.

The 2019 GDP map reveals the well-known large regional disparities. Dense clusters of high GDP are observed around major cities, such as São Paulo and Santiago in South America, Moscow in Russia, Beijing, Shanghai and Seoul in Asia, and Cairo and Johannesburg in Africa. Coastal areas, especially those along major maritime shipping routes, display higher GDP values compared to inland regions. For instance, the Asia-Europe route, connecting China's coastal regions through critical chokepoints like the Strait of Malacca and the Suez Canal, terminates in Europe via Gibraltar or the Cape of Good Hope in Africa, with some segments bypassing the Gulf of Guinea in West Africa. Key nodes along this route consistently exhibit high GDP levels. Similarly, the Atlantic coastal regions of Brazil along the South America-Europe route are economically prominent. Conversely, areas with challenging geographical features, such as the Himalayan Mountains, the Sahara and Arabian Deserts, central Australia's arid zones, and the Amazon rainforest, show low economic activity and the only GDP clusters are confined to urban centers and water-abundant locations.

The 2019 GDP per capita map also reveals significant regional disparities. Major cities exhibit higher GDP per capita compared to surrounding rural areas. In China, eastern coastal cities demonstrate substantially higher GDP per capita, and values decline progressively inland. Similarly, in Brazil, the northern regions have lower GDP per capita than the south, likely due to environmental and geographical challenges, limited infrastructure, or lower population density. Interestingly, some sparsely populated areas, for example some parts of Sub-Saharan Africa and Russia, exhibit high GDP per capita. This is often attributed to natural resource extraction, particularly oil and gas, which disproportionately

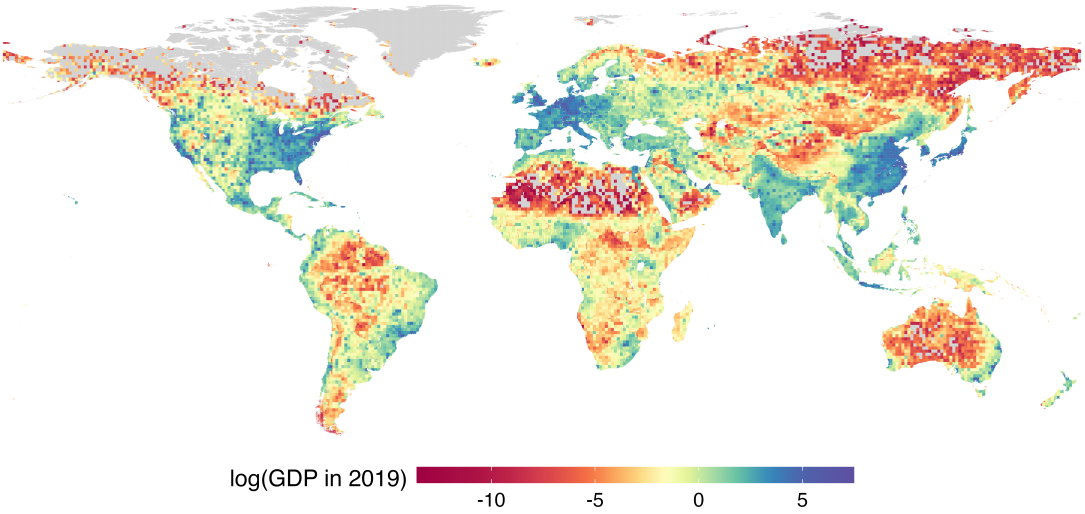
[13] We provide the standard deviation of cell predictions in the publicly available datasets on our website. A detailed discussion of these uncertainty measures is provided in Section 2.5 and Online Appendix Section 10.

contribute to GDP in these regions.

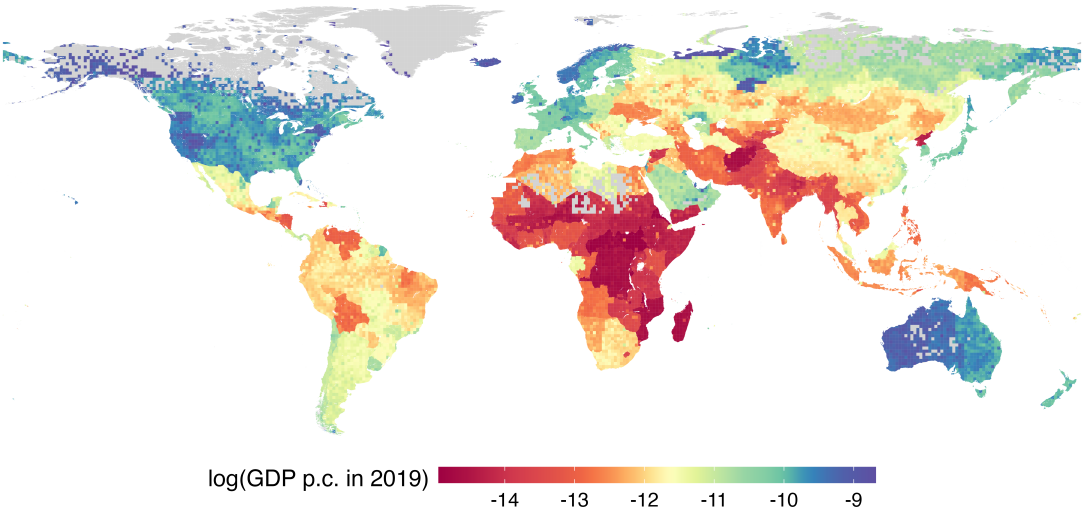
Panels (c) and (d) depict annual changes in GDP and GDP per capita. In sparsely populated areas, cells can show sharp changes and differ from adjacent cells, as seen in deep red and deep blue cells in Russia, Canada, Sub-Sahara Africa, and Australia. In contrast, densely populated areas exhibit more consistent changes, with neighboring cells showing similar levels of growth or decline. In 2019, positive growth in GDP and GDP per capita is concentrated in densely populated regions of North America, South Asia, Eastern Europe, and parts of Africa, while large negative growth is observed in South America, Europe, and Southern Africa. Notably, sharp cross-border contrasts are evident in Africa, whereas Asia, Europe, and North America display much less pronounced cross-border differences and more uniform regional growth.

2.2 The Relationship Between Population and GDP

A key aspect of assessing the usefulness of our estimates is the extent to which they capture factors beyond population. Figure 2 plots cell-level GDP against population count, in log-log scale, across three resolutions. It shows that our model reveals a concave relationship for very low-population-density areas and a largely convex relationship in most other regions. The concavity observed in sparsely populated areas may stem from congestion in production factors, like land, while the convexity aligns with standard agglomeration forces, where higher density yields productivity gains. This plot also shows a significant disparity: an average variance of approximately 7 log points of constant 2021 billion USD in cell GDP values with the same cell population. This substantial variation underscores the influence of factors beyond population in determining GDP.



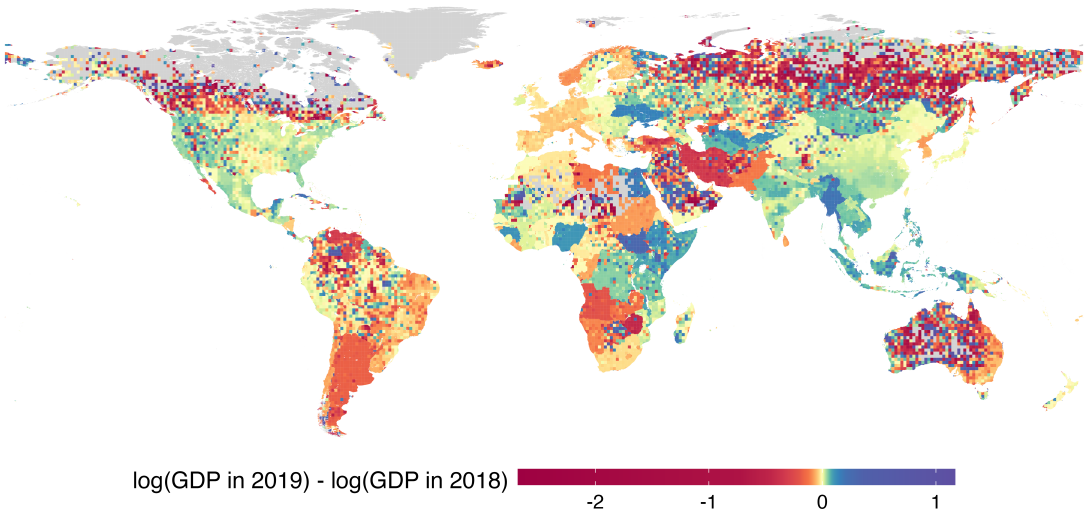
(a) Cell-level log GDP, 2019



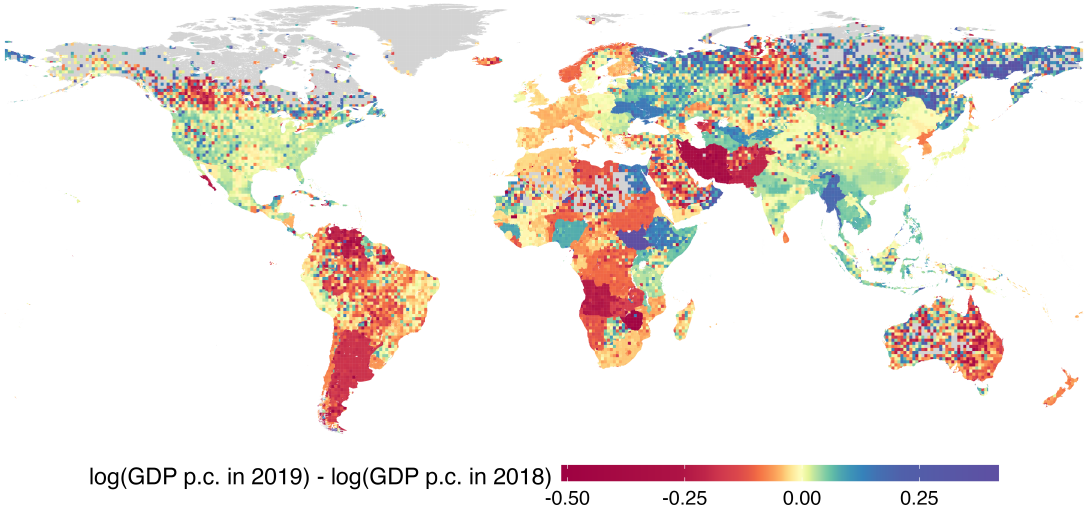
(b) Cell-level log GDP per capita, 2019

Figure 1: Map of Global 1-degree Cell log(GDP), log(GDP Per Capita) and Annual Growth Rates

Note: GDP is measured in constant 2021 billion USD. Panels (a) and (b) show 2019 levels; panels (c) and (d) show the year-over-year log change from 2018 to 2019. Grey cells are those with a population density of 0 people per km^2, which have been bottom-coded to a value of 0.



(c) Cell-level annual log change in GDP, 2018–2019



(d) Cell-level annual log change in GDP per capita, 2018–2019

(Figure 1, continued.)

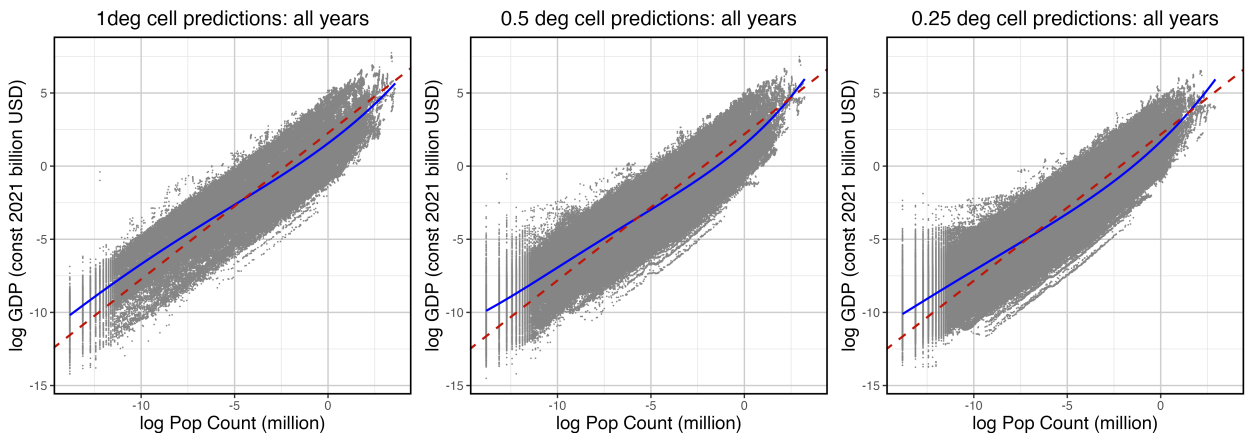


Figure 2: Cell Population Against Cell GDP

Note: For each resolution, the blue solid line represents the estimate from a fourth-degree polynomial regression, while the red dashed line represents the estimate from a linear regression. Both models are fitted using data from all years. For the 1-degree resolution, the polynomial regression is $y = 1.56 + 0.96x + 0.04x^2 + 0.003x^3 + 0.0001x^4$ and the linear regression is $y = x + 2.26$. For the 0.5-degree resolution, the polynomial regression is $y = 1.48 + 1.1x + 0.07x^2 + 0.006x^3 + 0.0002x^4$ and the linear regression is $y = x + 2.18$. For the 0.25-degree resolution, the polynomial regression is $y = 1.68 + 1.22x + 0.07x^2 + 0.004x^3 + 0.0001x^4$ and the linear regression is $y = x + 2.16$.

This characteristic of our predictions is key for researchers seeking to capture nuanced regional output differences and their drivers. For instance, if one aims to estimate local total factor productivity (TFP) across regions, analyze the impact of advanced technology adoption, variations in labor skill levels, measure the impacts of infrastructure improvements on regional welfare, or perform counterfactual analyses to identify optimal policies, a dataset that allocates GDP solely based on population may yield misleading conclusions. Similarly, if researchers seek to assess the effects of trade openness on regional economic resilience or evaluate how proximity to ports or major markets influences economic outcomes, GDP estimates based solely on population might obscure critical spatial patterns and underestimate the role of trade infrastructure.

2.3 Method Performance

As described in Section 1.2, we employ Group 5-Fold Cross-Validation, in which countries are partitioned into five groups. The model is trained on four groups and used to predict outcomes for the held-out fold of countries. Because the cells belonging to countries in the held-out fold are entirely excluded from the training set, they constitute genuine out-of-sample observations that are not contaminated by in-sample data points through spatial

(within-country) or temporal dependencies.[14] Table 1 presents the average results across these five folds, reporting metrics separately for developed, developing countries, and as a weighted mean of both groups. For each fold, the model is trained on a subset of the data (excluding one fold) using the optimal set of hyperparameters,[15] and then tested on the held-out fold. The R^2 values are calculated individually for developed, developing, and all countries, with the final R^2 values in the table representing the average across all five folds. The weighted value is computed as a weighted average of the developed and developing group values, with weights of 0.332 and 0.668, respectively, reflecting each group's share of total cells globally.

The R^2s of at least 0.94 shown in panel A of Table 1 indicate strong predictive power for GDP levels. The model at the 1-degree resolution shows slightly better performance compared to the other two resolutions, likely due to the lower potential for error with fewer cells. The model also shows slightly better performance for developed countries because of developed countries' availability of more reliable and detailed subnational data. Despite the coarser resolution of subnational data in developing countries (often exceeding 1-degree cells), we still include these regions.[16]

For year-over-year predictions of cell-level GDP changes, the models attain R^2 values of at least 73%. Again, more reliable and detailed regional data allow predictions for developed countries to outperform those for developing regions. When comparing across resolutions, the model at the 0.25-degree resolution shows the best performance, likely due to the method used to construct the true GDP data at this scale rather than an inherent advantage of the model at finer resolutions. Recall that true GDP values are calculated by assuming constant GDP per capita within each country and distributing GDP based on population, which

[14]While our main country-level holdout design is robust to within-country spatial and temporal spillovers by construction, we additionally assess how spatial autocorrelation affects performance metrics under a less stringent holdout. To this end, we implement a spatial block cross-validation exercise for the 1-degree model, in which grid cells within each country are partitioned into contiguous clusters of size S via k-means on cell centroids and randomly assigned to five folds. We vary S over {1, 2, 5, 10, 25, 50}, where larger S values impose greater spatial separation between training and held-out cells, and S equal to the country's cell count recovers our baseline country-level holdout. As reported in Online Appendix Section 8.2, Table 1 performance is remarkably stable across all cluster sizes: the overall R^2 for log GDP levels ranges from 97.93% to 98.05% (versus 97.95% for the baseline), and the within-country R^2 exhibits similar stability. For annual changes, R^2 varies modestly across cluster sizes but shows no systematic decline. Notably, even the S = 1 specification, in which individual cells are held out while neighboring cells remain in training, produces similar results. These results confirm that the model's predictive power does not rely on interpolating from spatially contiguous training cells. This design enables us to evaluate the out-of-sample predictive performance of our methodology on unseen countries. Under the optimal hyperparameters, this five-fold procedure yields five out-of-sample R^2 values.

[15]The optimal set of hyperparameters is the same as in Figures 1 and 2.

[16]In the Online Appendix Section 7.3, we compare models trained with and without data from developing countries.

is also used as a predictor in the model. At the 0.25-degree resolution, many counties' areas are larger than the corresponding grid cells, potentially aligning the training process with the predictor structure. This alignment may contribute to the higher performance metrics observed. However, this does not imply that population alone determines the model's predictions. As demonstrated in Section 2.2, the 0.25-degree model's predictions also reveal a convex relationship between GDP and population and local GDP still varies substantially at equal population levels.[17]

The R^2 values reported in Table 1, Panels A and B, represent overall R^2, computed by pooling all cells across countries. A concern is that these high R^2 values may partly reflect cross-country variation in GDP levels. To address this, we additionally compute the within-country R^2, obtained by first demeaning log GDP and log GDP changes and their predicted values by their respective country-year means, and then calculating R^2 on the resulting residuals. The within-country R^2 for log GDP levels exceeds 93% across all resolutions and country groups, only modestly below the overall R^2. For annual changes, the within-country R^2 ranges from 69% to 82%, again closely tracking the overall metrics. These results confirm that the model captures within-country variation rather than merely reflecting cross-country differences in GDP. Detailed results are reported in Online Appendix Section 8.1.

Panel C in Table 1 presents some of the key variables shared across the three resolutions and their corresponding importance scores. These scores are calculated by comparing the drop in R^2 derived from replacing each of the predictors used in estimating the model with its global mean. We also present scores using the correlation between the data and model's log difference of GDP share and population share ($Corr$).[18] The values are meaningful for comparing across predictors within a single model. The results reveal substantial contributions from diverse data sources in capturing economic activity. Urban population and population in other areas (neither urban nor cropland) rank among the top predictors, while nighttime lights (NTL) from cropland and cropland population contribute meaningfully, particularly at finer resolutions. Notably, CO_2 emissions from heavy industry emerge as increasingly important at the 0.5-degree and 0.25-degree resolutions, despite showing minimal importance

[17]The population variable used to construct the dependent variable differs from that entering the predictor set. Specifically, the left-hand side is constructed using population shares at the county level (or its regional equivalent), since the training data consist of observations with county-level or regional-level GDP. By contrast, the population shares used as predictors are measured at the state or country level. Hence, these are not the same variable, and so it does not mechanically inflate model performance. Nevertheless, to further assess the importance of this concern directly, we re-train the model at the 1-degree resolution, excluding total population share from the predictor set, and report the results in Online Appendix Section 7.4. The R^2 for log GDP levels declines modestly from 97.95% to 95.85% at the 1-degree resolution, while the R^2 for log GDP changes declines from 77.74% to 58.93%.

[18]See Online Appendix Section 3 for a detailed description of the method used for constructing these scores.

at the coarsest scale, suggesting that industrial activity indicators become more informative at finer resolutions. In Online Appendix Section 7.4, we discuss the robustness of our results to removing total population, CO_2 emissions, national GDP per capita, and all lags, respectively.

2.4 Model Performance During The COVID-19 Pandemic

To test our predictions during extraordinary, unprecedented, episodes, we use one of the most significant recent disruptions: the COVID-19 pandemic. China, which experienced considerable economic effects during the COVID-19 pandemic, has city-level GDP data that can be collected individually and was not included in our training set, making it an ideal out-of-sample test case. We collected city-level GDP data for seven major provinces in China: Guangdong, Henan, Hubei, Jiangsu, Shandong, Sichuan, and Zhejiang, and aggregated it into 1-degree cells. Figure 3 compares true and predicted values for pre-COVID years (2012 to 2019), the COVID year (2020), and post-COVID years (2021 to 2022). The predicted GDP values are obtained directly from our 1-degree model.

The model shows strong performance in predicting out-of-sample GDP levels and year-over-year growth rates during both the COVID-19 shock and the subsequent recovery. The R^2 are greater than 0.94 for levels and at least 0.52 for annual changes. For the COVID-19 and post-COVID years, the model's performance on annual changes exceeds 0.83, consistent with the cross-validated performance shown in Panels B of Table 1. The R^2 value of 0.52 for log changes in GDP prior to 2019 could be attributed to issues with data reliability. Only after 2019 did the Chinese central government implement a standardized approach for local governments to calculate regional GDP. This reform improved the comparability and accuracy of GDP data across regions.

Table 1: Cross-Validated Performance Metrics Across Spatial Resolutions

	1-degree Model	0.5-degree Model	0.25-degree Model
Panel A: R^2 of log GDP Level			
R^2 (Developed)	98.32%	97.45%	97.62%
R^2 (Developing)	95.94%	95.84%	94.63%
R^2 (All)	97.95%	97.16%	97.04%
Weighted R^2	96.60%	96.29%	95.46%
Panel B: R^2 of $\log(GDP_t) - \log(GDP_{t-1})$			
R^2 (Developed)	73.55%	74.47%	84.05%
R^2 (Developing)	79.37%	76.60%	81.03%
R^2 (All)	76.20%	73.51%	81.88%
Weighted R^2	77.74%	76.00%	81.87%

Panel C: Variable Importance Scores

	R^2	*Corr*	R^2	*Corr*	R^2	*Corr*
Population (urban)	21.29	0.1933	41.02	0.1999	27.13	0.1682
Lag population (urban)	32.21	0.2020	40.46	0.1969	11.71	0.1569
Population (other)	20.23	0.1815	19.61	0.1805	18.72	0.1599
Lag CO2 non-org (heavy industry)	0.05	0.0099	16.12	0.1810	10.25	0.1580
CO2 non-org (heavy industry)	0.02	0.0028	14.77	0.1795	11.66	0.1589
Population (cropland)	1.30	0.1038	8.34	0.1608	8.17	0.1489
Lag population (cropland)	0.07	0.0165	7.09	0.1567	5.48	0.1416
Lag population (other)	4.47	0.1433	6.96	0.1559	6.52	0.1464
NTL (cropland)	3.55	0.1365	6.01	0.1534	4.79	0.1429
Lag NTL (urban)	0.18	0.0528	4.30	0.1433	1.58	0.1068
NTL (urban)	0.37	0.1021	3.48	0.1363	1.95	0.1152
Lag NTL (cropland)	2.70	0.1278	2.92	0.1292	1.81	0.1115
Lag CO2 bio (manuf. combust.)	0.03	0.0055	2.08	0.1253	0.97	0.1029
NTL other (snow-free period)	0.02	0.0096	2.02	0.1177	0.26	0.0418
CO2 bio (manuf. combust.)	0.07	0.0196	0.98	0.0995	0.34	0.0581
Population	0.76	0.1226	0.69	-0.1073	0.97	-0.1295
CO2 non-org (manuf. combust.)	0.05	0.0184	0.29	0.0429	0.08	0.0144
Lag CO2 non-org (manuf. combust.)	0.05	0.0202	0.21	0.0356	0.10	0.0189
Lag cropland	0.14	0.0323	0.04	0.0076	0.01	0.0025
Cropland	0.11	0.0241	0.04	0.0089	0.01	0.0030

Note: Panels A and B report cross-validated overall R^2 metrics. Panel C reports variable importance scores: drop in within-country-year R^2 (expressed as decimals) and *Corr* when each variable is replaced with its global mean. Variables shown are the top 20 by their maximum R^2 importance score across all degree levels. Please refer to Section 3 of the Online Appendix for a detailed discussion of the methodology used for constructing these scores.

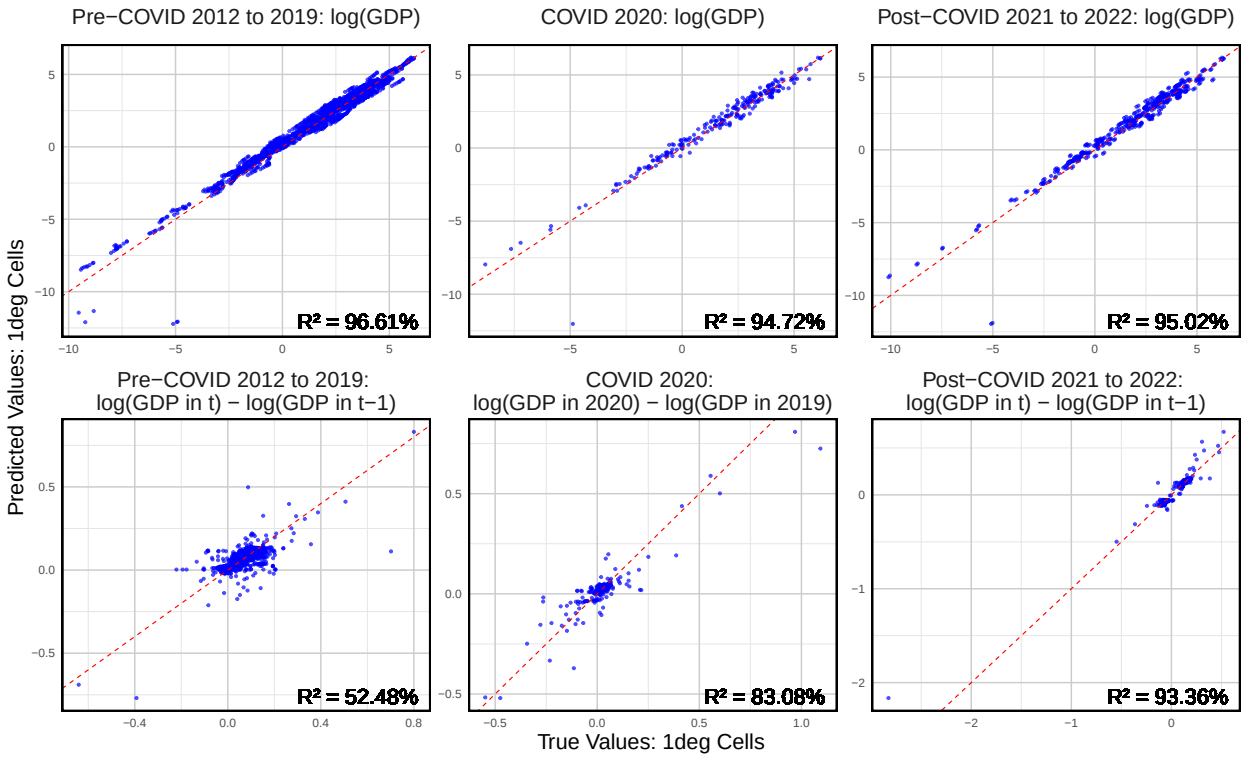


Figure 3: Model Predictions Against Actual Values in Billion Constant 2021 USD for Seven Leading Provinces in China

Note: The red dashed line represents the 45-degree line. Cells with a GDP value of zero are omitted to enable the calculation of logarithmic values. Only two out of 2618 data points are excluded due to zero GDP values.

To further demonstrate that the model's robust performance on COVID-19 affected data does not arise from prior exposure to COVID-19 years in the training dataset, we repeated the analysis by training the model on all available countries excluding China (as in Figure 3), but also exclude data from 2020 to 2022. We then tested the model predictions on the years 2020 to 2022. The results, which are shown in Online Appendix Section 6, maintain comparable R^2 values and high predictive accuracy. This demonstrates that the model's good performance in these tests does not rely on training on the COVID-19 experience in other countries.

Wuhan, one of the COVID-19 affected cities, can provide a useful case study for assessing the model's accuracy. Figure 4 compares the actual annual change in log GDP with our estimates for Wuhan's main economic activity hub from 2012 to 2022. Our model effectively captures the city's GDP growth fluctuations, successfully identifying periods of slower growth in 2016 and years of rapid expansion in 2017 and 2018. It also identifies the large economic disruptions in 2020 caused by the COVID-19 pandemic and the subsequent recovery in 2021

(only to revert back again in 2022). These findings demonstrate that the statistical model can successfully identify fluctuations in GDP growth rates generated by unique unanticipated shocks.

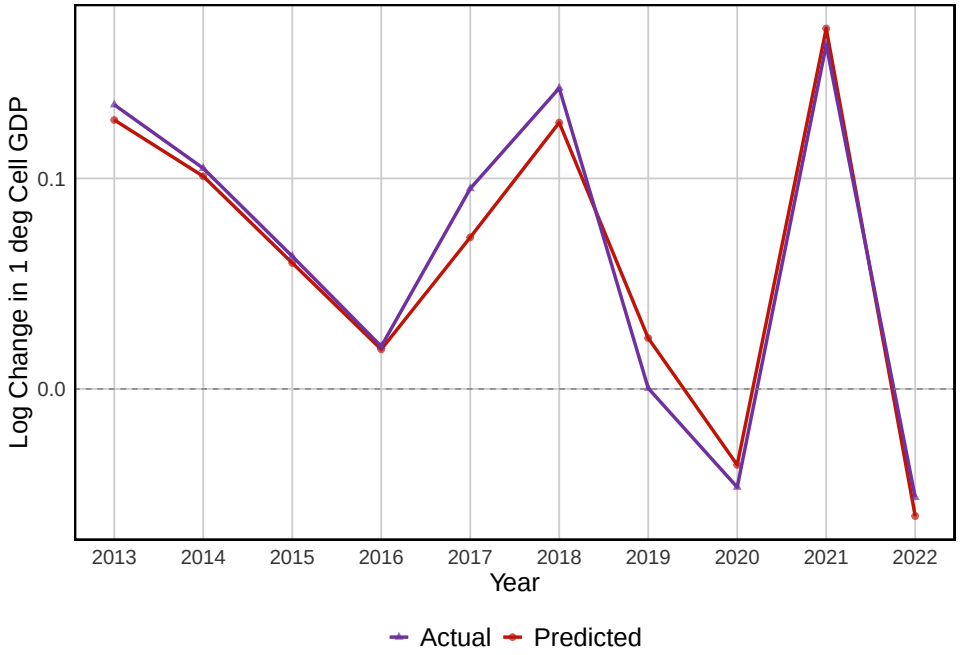


Figure 4: Model Predictions Against Actual Values of Annual Change in Log GDP for Cell 21535: Main Economic Activity Hub of City Wuhan

2.5 Prediction Uncertainty

Our random forest model provides a natural measure of uncertainty for each cell's predictions. The random forest comprises many individual trees, each of which generates a prediction of the GDP share for every cell. For each tree, we rescale the predicted cell shares within each year-country (or state, when available) pair to sum to one and then compute cell-level GDP by multiplying these shares by aggregate country-year (or state-year) GDP. Repeating this procedure across all trees yields a distribution of log GDP predictions for each cell, from which we compute the standard deviation. A larger standard deviation indicates greater disagreement among trees and, correspondingly, greater model uncertainty regarding that cell's GDP predictions. Figure 5 maps this measure for the 0.25-degree model in 2019. Naturally, uncertainty is high in sparsely populated areas such as the Sahara, Central Asia, and interior parts of Australia, where predictor values are near zero and provide little basis for distinguishing one cell from another. Online Appendix Section 10 reports the corresponding maps for the 1-degree and 0.5-degree models and a further breakdown by in-sample versus

out-of-sample countries and by development status.

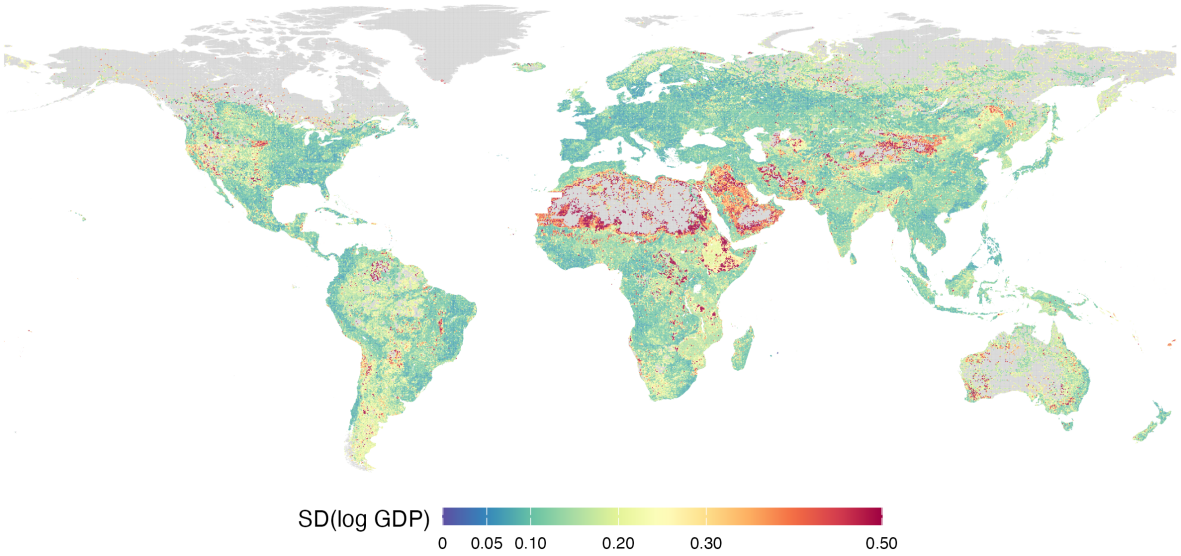


Figure 5: Standard Deviation of log(GDP) Across Random Forest Trees (2019, 0.25-degree)

Note: Cell-level SD of log(GDP) across all trees in the random forest ensemble. Higher values indicate greater prediction uncertainty. Values above 0.5 are top-coded for visualization.

Figure 6 presents the cross-cell distribution of this uncertainty measure. The distribution is right-skewed, consistent with the geographic pattern displayed in Figure 5: the majority of cells exhibit low uncertainty, while a tail of sparsely populated cells accounts for the upper end of the distribution. The median cell has a standard deviation of 0.15 in log GDP, corresponding to approximately 15% proportional uncertainty in levels, and 73% of cells have a standard deviation below 0.20. Online Appendix Section 10 reports a combined kernel density plot overlaying the cross-cell distributions of this uncertainty measure across all three resolutions. We also include the standard deviation of predictions for each cell in the publicly available datasets on our website.[19]

[19]See Online Appendix Section 11 for maps and distributions of the log prediction error for all three resolutions.

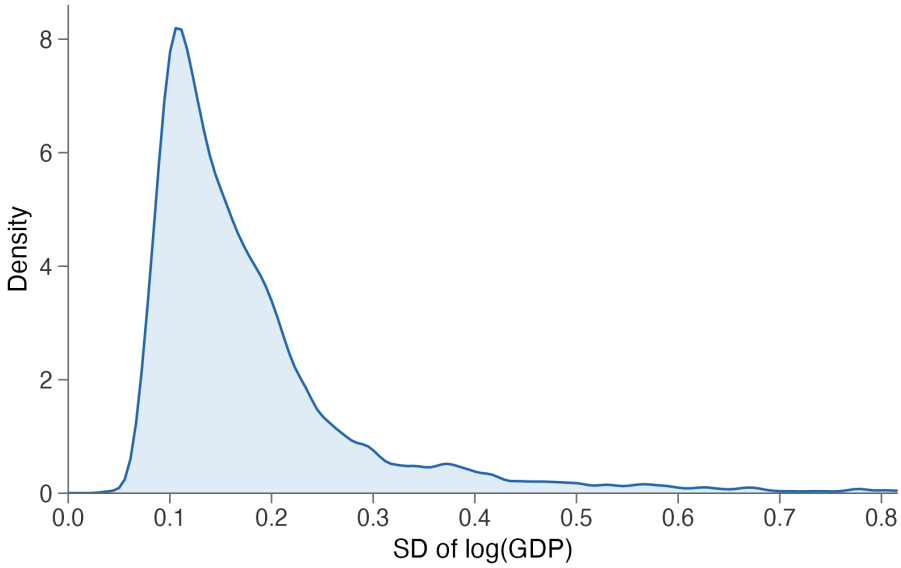


Figure 6: Distribution of SD of log(GDP) Across Cells (2019, 0.25-degree)

Note: Kernel density estimate of the cell-level SD of log(GDP) across random forest trees, based on 186,732 cells in 2019. All cells (i.e., both in-sample and out-of-sample) with positive predicted GDP are included. Estimation uses a Gaussian kernel with bandwidth selected by Silverman's rule of thumb.

3 Discussion and Robustness

A discussion of other potential concerns and the robustness of our estimates is in order. First, one might be concerned that allocating aggregate output within larger administrative units introduces artificial discontinuities at borders. Our methodology imposes no additional adjustments or constraints on cells near administrative boundaries, so we can directly examine whether border cells exhibit larger prediction errors than interior cells. We find that border cells, on average, have an absolute log GDP prediction error that is 0.0125 log points smaller than that of interior cells, conditional on population share, cell area, and country fixed effects, with an R^2 of 98.12% compared to 97.53% for interior cells at the 1-degree resolution. These results indicate that our model does not generate inflated errors or artificial discontinuities at administrative borders. See Online Appendix Section 9.1 for details.

Second, because our model incorporates additional predictors beyond population, one might also be concerned that the CO_2-related predictors mechanically inflate predicted GDP along major transportation corridors, or that predicted GDP is distorted in cropland-heavy cells, where changes in agricultural output driven by new crop adoption may not be fully captured by photosynthesis or reflectance indices. To investigate the first concern, we define corridor intensity as the ratio of a cell's transportation CO_2 emissions share to its population

share within the same country, and compare model performance across this measure. The top 10% most transport-intensive cells show out-of-sample R^2 values that are only 0.38 to 2.76 percentage points lower than those of the remaining 90%. A regression of the signed log prediction error on corridor intensity (controlling for population share and country fixed effects) yields a positive and statistically significant coefficient at all three resolutions, but an additional over-prediction of only around 0.01% per unit of corridor intensity. An analogous exercise using cropland share yields R^2 differences of $+0.05$, -0.14, and -0.98 percentage points at the 1-, 0.5-, and 0.25-degree resolutions, respectively, and a statistically insignificant coefficient in the corresponding error regression. Taken together, these results indicate that both the corridor-driven mechanical bias and the agriculture-related concern are detectable but quantitatively modest; detailed results are reported in Online Appendix Sections 9.2, 9.3, and 9.4.

Third, although our model predicts GDP shares, the hyperparameters are tuned to maximize out-of-sample performance on log GDP changes. These share-based predictions of local GDP and GDP changes anchor cell-level estimates to observed regional GDP totals and generate more reliable local data. However, predictor shares (e.g., NTL shares) change interdependently across cells, and because predicted GDP shares inevitably contain some prediction error, a cell with no real economic change may still show apparent GDP changes due to these errors. As a result, downstream researchers using the data for causal analysis, spatial spillover estimation, or the study of localized economic shocks should be mindful that predicted changes in a cell's GDP may partly reflect prediction error rather than genuine local economic dynamics. To gauge the importance of this concern, we conducted an exercise in which predicted GDP shares are not rescaled to sum to one, with all other steps held constant, to isolate the effect of the rescaling step. The results, reported in Online Appendix Section 8.3, show that out-of-sample performance is only very mildly affected at all three resolutions.

Finally, although our random forest prediction model includes lagged variables as predictors, it targets GDP levels rather than year-over-year changes. One may be concerned that this limits the usefulness for applications that depend on temporal variation. Nonetheless, we tune hyperparameters to maximize the R^2 on log GDP changes, so predictive accuracy on changes is built into the model selection procedure. Targeting levels rather than changes serves our main objective of producing sub-national GDP estimates for countries with only national data. Of course, a change-targeted specification can recover year-over-year variation but cannot generate a level for cells without an initial local observation, whereas our approach yields both levels and changes wherever national GDP is available. In addition, directly modeling changes would require using first differences of noisy predictors such as

population, nighttime lights, and CO_2 emissions, each of which is itself the output of a remote-sensing or some data construction procedure; differencing these series amplifies the underlying measurement error, so their levels are substantially more informative than their changes. We emphasize, however, that our high R^2 on log GDP changes reflects the model's ability to capture the overall distribution of changes across cell-years, not a guarantee of small absolute error for any individual cell. Researchers whose analyses rely heavily on within-cell temporal variation should therefore interpret our predictions with caution at the cell-year level.

4 Conclusion

In this paper, we present a new method to predict out-of-sample local GDP based on predictor variables—such as remote sensing data, population data, and emissions data. By training the model on countries with detailed and accurate subnational data, we extend local GDP estimates to regions lacking granular economic data. The predictor data used are updated annually, allowing our global local GDP estimates to also be updated yearly and extend beyond the current 2012–2022 range. All data are accessible for download at `http://bfidatastudio.org/gdp`.

One immediate application of this dataset is its direct use as an outcome variable in various types of research and analysis. This is particularly relevant in development economics, where researchers often rely on nighttime light intensity as a proxy for economic development or income indicators. Our dataset, however, offers more precise subnational data to assess regional disparities and the development of low-income areas. For example, it could refine insights in studies like Storeygard (2016)'s analysis of economic growth in Sub-Saharan African cities, investigations into the effects of sanctions on North Korea (Lee, 2018), or analyses of the impact of agroclimatic similarity on migrant productivity in Indonesia (Bazzi et al., 2016).

This dataset also has significant implications for research using structural spatial models. These models often need observed local GDP per capita to estimate local productivity parameters through inversion procedures. Or, when applying exact-hat algebra, structural models may also require reliable income data to assess shock impacts. Compared to G-Econ, which relies heavily on population-based estimates or GDP data derived only from nighttime lights, our dataset can enhance the accuracy of counterfactual predictions and provide additional data points and years for model estimation. This improvement benefits studies analyzing the heterogeneous effects of global warming across regions (Cruz and Rossi-Hansberg, 2024), assessing natural disaster impacts (Desmet et al., 2021), evaluating the effects of easing mi-

gration restrictions (Desmet, Nagy and Rossi-Hansberg, 2018), identifying optimal transport networks (Fajgelbaum and Schaal, 2020), and analyzing the welfare effects of transportation infrastructure improvements (Allen and Arkolakis, 2022) beyond the United States.

A final word of caution is in order. While our dataset effectively captures the within-country and within-state distributions of log GDP levels and annual changes, predictions for some individual cells may still contain substantial error, particularly in sparsely populated regions or areas with limited training signal. Accordingly, users conducting cell-level analyses, such as studies of highly localized economic shocks, narrowly defined geographic boundaries, or fine-grained causal inference relying on individual cell values, should exercise particular caution and, where possible, incorporate the cell-level uncertainty estimates provided alongside the GDP predictions. For applications involving spatial aggregation or comparisons across larger regions, prediction errors tend to average out, and the dataset can be used with greater confidence. Nonetheless, we believe this new dataset remains a valuable and robust resource for economic research.

References

Allen, Treb, and Costas Arkolakis. 2022. "The Welfare Effects of Transportation Infrastructure Improvements." *The Review of Economic Studies*, 89(6): 2911–2957.

Bazzi, Samuel, Arya Gaduh, Alexander D Rothenberg, and Maisy Wong. 2016. "Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia." *American Economic Review*, 106(9): 2658–2698.

Breiman, Leo. 2001. "Random Forests." *Machine Learning*, 45: 5–32.

Bright, E., A. Rose, M. Urban, K. Sims, J. McKee, A. Reith, et al. 2012–2021. "LandScan Global Population Dataset." Oak Ridge National Laboratory. https://landscan.ornl.gov/ (accessed Jan 1, 2024).

Chen, Jiandong, Ming Gao, Shulei Cheng, Wenxuan Hou, Malin Song, Xin Liu, and Yu Liu. 2022. "Global 1 km× 1 km Gridded Revised Real Gross Domestic Product and Electricity Consumption During 1992-2019 Based on Calibrated Nighttime Light Data." *Scientific Data*, 9(1): 202.

Cruz, José-Luis, and Esteban Rossi-Hansberg. 2024. "The Economic Geography of Global Warming." *Review of Economic Studies*, 91(2): 899–939.

Desmet, Klaus, Dávid Krisztián Nagy, and Esteban Rossi-Hansberg. 2018. "The Geography of Development." *Journal of Political Economy*, 126(3): 903–983.

Desmet, Klaus, Robert E. Kopp, Scott A. Kulp, Dávid Krisztián Nagy, Michael Oppenheimer, Esteban Rossi-Hansberg, and Benjamin H. Strauss. 2021. "Evaluating the Economic Cost of Coastal Flooding." *American Economic Journal: Macroeconomics*, 13(2): 444–86.

Emissions Database for Global Atmospheric Research (EDGAR). 2012-2021. "Global Greenhouse Gas Emissions, Version 8.0, 2012-2021." European Commission, Joint Research Centre (JRC), and International Energy Agency (IEA). https://edgar.jrc.ec.europa.eu/ (accessed Jan 1, 2024).

Fajgelbaum, Pablo D, and Edouard Schaal. 2020. "Optimal Transport Networks in Spatial Equilibrium." *Econometrica*, 88(4): 1411–1452.

Friedl, M., and D. Sulla-Menashe. 2022. "MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061 [Data Set]." NASA EOSDIS Land Processes Distributed

Active Archive Center. https://doi.org/10.5067/MODIS/MCD12Q1.061 (accessed Jan 1, 2024).

Global Gas Flaring Data. 2012-2023. "Flare Volume Estimates by Individual Flare Location." National Oceanic and Atmospheric Administration, Payne Institute at the Colorado School of Mines, World Bank, Global Flaring and Methane Reduction Partnership (GFMR). https://www.worldbank.org/en/programs/gasflaringreduction/global-flaring-data (accessed Jan 1, 2024).

Henderson, J Vernon, Adam Storeygard, and David N Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review*, 102(2): 994–1028.

Henderson, Vernon, Tim Squires, and David N. Weil. 2018. "The Global Spatial Distribution of Economic Activity: Nature, History, and the Role of Trade." *Quarterly Journal of Economics*, 133(1): 357–406.

Keola, Souknilanh, Magnus Andersson, and Ola Hall. 2015. "Monitoring Economic Development from Space: Using Nighttime Light and Land Cover Data to Measure Economic Growth." *World Development*, 66: 322–334.

Khachiyan, Arman, Anthony Thomas, Huye Zhou, Gordon Hanson, Alex Cloninger, Tajana Rosing, and Amit K Khandelwal. 2022. "Using Neural Networks to Predict Microspatial Economic Growth." *American Economic Review: Insights*, 4(4): 491–506.

Lee, Yong Suk. 2018. "International Isolation and Regional Inequality: Evidence from Sanctions on North Korea." *Journal of Urban Economics*, 103: 34–51.

Nordhaus, William D. 2006. "Geography and Macroeconomics: New Data and New Findings." *Proceedings of the National Academy of Sciences*, 103(10): 3510–3517.

Nunn, Nathan, and Diego Puga. 2012. "Ruggedness: The Blessing of Bad Geography in Africa." *Review of Economics and Statistics*, 94(1): 20–36.

Román, M. O., Z. Wang, Q. Sun, V. Kalb, S. D. Miller, A. Molthan, L. Schultz, J. Bell, E. C. Stokes, B. Pandey, K. C. Seto, et al. 2018. "NASA's Black Marble Nighttime Lights Product Suite." *Remote Sensing of Environment*, 210: 113–143.

Running, S., and M. Zhao. 2021. "MODIS/Terra Net Primary Production Gap-Filled Yearly L4 Global 500m SIN Grid V061 [Data Set]." NASA EOSDIS Land Processes

Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD17A3HGF.061 (accessed Jan 1, 2024).

Storeygard, Adam. 2016. "Farther on Down the Road: Transport Costs, Trade and Urban Growth in Sub-Saharan Africa." *The Review of Economic Studies*, 83(3): 1263–1295.

Vogel, Kathryn Baragwanath, Gordon H Hanson, Amit Khandelwal, Chen Liu, and Hogeun Park. 2024. "Using Satellite Imagery to Detect the Impacts of New Highways: An Application to India." *NBER Working Paper No. 32047*.

Wenz, Leonie, Robert Devon Carr, Noah Kögel, Maximilian Kotz, and Matthias Kalkuhl. 2023. "DOSE–Global Data Set of Reported Sub-national Economic Output." *Scientific Data*, 10(1): 425.