



This is an electronic reprint of the original article. This reprint may differ from the original in pagination and typographic detail.

Kummu, Matti; Kosonen, Maria; Sayyar, Sina Masoumzadeh

Downscaled gridded global dataset for gross domestic product (GDP) per capita PPP over 1990–2022

Published in: Scientific Data

*DOI:* 10.1038/s41597-025-04487-x

Published: 01/12/2025

Document Version Publisher's PDF, also known as Version of record

Published under the following license: CC BY

Please cite the original version:

Kummu, M., Kosonen, M., & Sayyar, S. M. (2025). Downscaled gridded global dataset for gross domestic product (GDP) per capita PPP over 1990–2022. *Scientific Data*, *12*, Article 178. https://doi.org/10.1038/s41597-025-04487-x

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

# scientific data

Check for updates

## **OPEN** Downscaled gridded global dataset for gross domestic product (GDP) DATA DESCRIPTOR per capita PPP over 1990-2022

Matti Kummu 🗅 🖾, Maria Kosonen 🗈 & Sina Masoumzadeh Sayyar 🕩

We present a comprehensive gridded GDP per capita dataset downscaled to the admin 2 level (43,501 units) covering 1990–2022. It updates existing outdated datasets, which use reported subnational data only up to 2010. Our dataset, which is based on reported subnational GDP per capita data from 89 countries and 2,708 administrative units, employs various novel methods for extrapolation and downscaling. Downscaling with machine learning algorithms showed high performance (R<sup>2</sup> = 0.79 for cross-validation, R<sup>2</sup> = 0.80 for the test dataset) and accuracy against reported datasets (Pearson R = 0.88). The dataset includes reported and downscaled annual data (1990–2022) for three administrative levels: 0 (national; reported data for 237 administrative units), 1 (provincial; reported data for 2.708 administrative units for 89 countries), and 2 (municipality: downscaled data for 43,501 administrative units). The dataset has a higher spatial resolution and wider temporal range than the existing data do and will thus contribute to global or regional spatial analyses such as socioenvironmental modelling and economic resilience evaluation. The data are available at https://doi. org/10.5281/zenodo.10976733.

### **Background & Summary**

The rapid increase in the availability of various global earth system and climate datasets over the past few decades<sup>1</sup> has made it possible to conduct different risk and vulnerability assessments at fine spatial scales globally<sup>2</sup>. To understand the effects on human life, these earth system and climate data should be analysed together with long-term continuous socioeconomic datasets<sup>3</sup>. However, traditionally, socioeconomic data are provided on a national level, creating a mismatch in scales and averaging out within-country heterogeneity. This, in turn, might lead to high inaccuracies in the results. Over the past decade, an increasing number of subnational or gridded datasets for various socioeconomic variables have been published, including data on migration patterns<sup>4</sup>, urbanisation<sup>5</sup>, gross national income (GNI) per capita and income inequality (Gini coefficient)<sup>6</sup>, and the human development index (HDI)<sup>7,8</sup>.

One of the most commonly used indicators for economic development is gross domestic product (GDP). While there are global subnational GDP per capita (or gross regional product (GRP) per capita) datasets, they are rather old—they end in 2015 with the latest subnational data from 2010 (ref. 8) or for only one year, 2005 (ref. <sup>9</sup>)—or they do not cover data for all countries and each year (DOSE v2 dataset<sup>10</sup>) or both<sup>11</sup>. Thus, there is no harmonised and gap-filled global subnational or gridded GDP per capita dataset that spans over recent decades. Here, we constructed a global harmonised and gap-filled GDP per capita dataset (PPP) for 1990-2022 downscaled to the admin 2 level (n = 43,501 administrative units). We provide the data both as gridded (5 arc-min resolution) and as polygon data (with administrative area names) for three administrative levels: the admin 0 level (i.e., national), the admin 1 level (where available; otherwise, the admin 0 level) (i.e., provincial), and the admin 2 level (i.e., municipality). Furthermore, we combined the downscaled product with population count data and estimated total GDP (PPP) for each grid cell at three resolutions: 30 arc-sec (for every 5 years), 5 arc-min (for each year), and 30 arc-min (for each year).

This dataset builds on the existing subnational GDP per capita (PPP) dataset by Kummu et al.<sup>8</sup>. However, here, we introduce several updates in terms of input data as well as methodological developments. First, while the subnational data used for Kummu et al.<sup>8</sup> are available only until 2010 (the dataset spans 2015, but the admin 1 level distribution of GDP per capita is based on up to 2010, while national data are available until 2015), here, the admin 1 level data expand until 2021, and the admin 0 level data expand until 2022. We also found data for

Water and Development Research Group, Aalto University, Espoo, Finland. Ze-mail: matti.kummu@aalto.fi

Study	admin 0 level areas	admin 1 level areas	admin 2 level areas	Temporal coverage
This study	Reported data for 237 countries	Reported data for 2708 admin units, for 89 countries	Downscaled to 43,501 admin units; for 237 countries	1990–2022 (reported subnational data 1990–2021)
Admin 0 level existing datasets				
World Bank	Reported data for 194 countries	NA	NA	1960-2023
IMF (International Monetary Fund)	Reported data for 195 countries	NA	NA	1980-2023
Admin 1 level existing datasets				
Kummu et al. <sup>8</sup>	Reported data for 231 countries	Reported data for 1549 admin units, for 83 countries	NA	1990–2015 (subnational data 1990–2010)
DOSE v2 <sup>10</sup>	NA	Reported data for 1,661 subnational units, across 83 countries (not full coverage for all countries)	NA	1953–2020
Wang and Sun <sup>9</sup>	Reported data for 197 countries	Reported data for ca. 800 subnational units, for 48 countries	NA, but downscaled to gridded level	Year 2005

Table 1. Comparison of existing global subnational GDP per capita products with this study.

.....

both the admin 0 and 1 level data for more countries—covering the admin 1 level data for 89 countries (2,708 subnational units) and the admin 0 level data for 237 countries—outperforming the existing datasets, as shown in Table 1. Finally, while in Kummu *et al.*<sup>8</sup>, the finest scale was the admin 1 level, here, we downscaled the data to the admin 2 level. Wang and Sun<sup>9</sup> also downscaled their data (to the gridded level), but their reported subnational data are based on rather few countries (Table 1), of which most are OECD countries, and only one year (2005).

We developed methods for multiple fronts, as briefly described below (see Methods for a more detailed explanation):

- We used a novel way of extrapolating missing data at the admin 0 level (the same method is also used in another article of ours, Chrisendo *et al.*<sup>12</sup>). While previously regional trends were used for extrapolation<sup>8</sup>, here, we fitted the available data of a country in question with countries with full or nearly full data extent via linear regression. We then selected the geographically closest country from among the best seven fits and used that regression model to estimate the missing years from the beginning and/or end of the time series.
- At the admin 1 level, we first calculated the 'admin 1 level/admin 0 level' ratio for the observed values (here called the subnational GDP ratio). The admin 0 level value, for each year for which data were available in a country in question, was estimated as the population weighted mean. Only after that we interpolated the values between years with reported data (for the missing data at the beginning and end of the study period, the last reported subnational GDP ratio was used). In this way, we were able to combine data from different sources, and it did not matter whether the original data were in local currency or USD. Only at the very end we multiplied the subnational GDP ratio with the gap-filled admin 0 level data of the country in question.
- We developed a novel way to downscale the admin 1 level data (admin 1 level where available, and elsewhere admin 0 level data) to the admin 2 level data using independent datasets (such as urbanisation level, travel time to closest city and income inequality) to train a set of machine learning models in which we used boosted ensembled trees with very good model performance ( $R^2 = 0.79$  for cross-validation and  $R^2 = 0.80$  for the testing set of the data).

Finally, while existing GDP per capita datasets provide data only at the admin 1 level<sup>8</sup> or admin 1 level and gridded scale<sup>9</sup>, here, we provide data for three levels: the admin 0, admin 1, and admin 2 levels. For all these levels, both gridded and polygon data are shared.

Our dataset thus provides more up-to-date and finer-scale global GDP per capita PPP (and total GDP PPP) data with longer temporal resolutions than the existing datasets do. It can be used for various global or regional analyses, covering topics such as climate change impacts and associated risks<sup>13</sup>, exposure to natural hazards<sup>14,15</sup>, urban development and urbanisation patterns<sup>16</sup>, biodiversity conservation and species invasion<sup>17</sup>, economic growth, inequality<sup>18</sup>, and sustainable development<sup>19</sup>. Furthermore, our data – gaps filled over time and down-scaled to the admin 2 level – might be particularly useful in data-scarce regions where high-resolution data are not available.

#### **Methods**

We first collected admin 0 level (i.e., national) GDP per capita at purchasing power parity (PPP) data from various sources, and the gaps for missing years were filled via linear interpolation and a novel extrapolation method (see Section 2.1). We then collected admin 1 level (i.e., provincial) data from sources such as the OECD, Eurostat, and national censuses. We calculated the ratio between the admin 1 level value and the admin 0 level value (estimated from the subnational values with population data as a weight) and calculated the subnational GDP ratio (ratio over the admin 1 level value and calculated the national value). We then interpolated the subnational GDP ratio to fill the gaps between years with reported values and used the latest subnational GDP ratio for the tail and leading missing values (see Section 2.2). These subnational GDP ratios were then used with the reported admin 0 level data to estimate the GDP per capita (PPP) value at the admin 1 level. Finally, we used the subnational GDP ratio to train a here developed downscaling method based on machine learning algorithms to



Fig. 1 Flow diagram of the study.

downscale the ratio from the admin 1 level to the admin 2 level (see Section 2.3). The overall workflow is illustrated in Fig. 1, and a more detailed explanation is given below. We used R (version 4.3.2) to conduct the study, except for downscaling, which was performed using MATLAB (version 2024a).

**Gap-filled national (admin0 level) dataset.** We first collected admin 0 level (national) GDP per capita (PPP) data for 1985–2022. We combined several databases, including the World Bank<sup>20</sup>, the International Monetary Fund (IMF)<sup>21</sup>, and the data based on the Central Intelligent Agency (CIA), accessed through IndexMundi<sup>22</sup>, with the preference of this order (i.e., if no data were available from the World Bank, then we used the IMF, and if they were not available, we turned to the CIA). We were able to collect data for 237 nations or sovereign states, making this the most comprehensive national GDP per capita PPP dataset available. The data are in 2017 international USD.

Countries had missing data for varying numbers of years in these datasets. To fill these gaps, we used a novel methodology for filling in the missing values (also used in another article by the lead author, i.e., Chrisendo *et al.*<sup>12</sup>), considerably improving the existing methods by, for example, Kummu *et al.*<sup>8</sup>, who used regional trends for extrapolation. In the updated method, we first used linear interpolation to fill the gaps between reported values. This was performed with the na.approx function in R (under the zoo package<sup>23</sup>). In general, there were not many 'holes' between the reported years. However, a larger issue was missing values at the beginning or at the end of the time series of a country. We thus developed a multistep extrapolation method to fill in these missing values:

- We first divided the countries into the following groups, based on the data coverage over the study period 1990–2022: a) full data extent, b) nearly full data extent (max three years missing from beginning or end), c) limited data extent (more than three years missing from either end but more than 5 data points), and d) very limited data extent (fewer than 5 data points).
- 2. We went through the nearly full data extent data country by country and extrapolated those using the full



**Fig. 2** Subnational data availability over time and space. (**a**) Number of reported years (i.e., if there were data for 2000, 2005, and 2010, then the number of reported years would be 3) for each administrative area; (**b**) range of reported years (i.e., if there were data for 2000, 2005, and 2010, then the range would be 11 years) for each administrative area; and (**c**) mean interval of reported years (i.e., if there were data for 2000, 2005, and 2010, then the mean interval would be 5 years).

data extent countries. This was done by constructing a linear model (lm) between a country with missing values (targetCountry) and each country with a full extent. Next, we filtered out seven countries with the best fits based on R<sup>2</sup> and from these, we chose geographically the closest country (bestClosestCountry) to the targetCountry using the centroids of each country and the distance between them. Im was then used to estimate the full time series (ImTimeseries) for a targetCountry using the data from bestClosestCountry as an input. Finally, we used the first and last reported values of the targetCountry and the corresponding values from the ImTimeseries to calculate the ratio. We used these ratios to scale the ImTimeseries to fill in the missing values from the beginning (ratio over the first nonmissing value) and end (ratio over the last nonmissing value) of the study period.

- 3. We combined the full dataset with the filled near-full dataset to create a combined full\_nearlyFull dataset.
- 4. We then filled in the missing data points for the limited data extent using the full\_nearlyFull dataset and the same method as in step 2.
- 5. We filled the countries with fewer than 5 data points by first identifying the closest country within the full\_nearlyFull dataset for each country in this group (closestCountry). We then scaled the leading and trailing missing values based on the trajectory of the closestCountry in a similar way to point step 2, i.e., we calculated the ratio between the first nonmissing value and the corresponding closestCountry value. We used this to estimate the leading missing values and the ratio between the last nonmissing value and the corresponding closestCountry value. These ratios were then combined with full data series from the closestCountry to estimate the missing values.
- 6. Finally, we combined the full and filled datasets to create a complete dataset.

**Subnational-level dataset.** We collected reported subnational-level GDP per capita (PPP) datasets from various sources. We started from the historical subnational database of Gennaioli *et al.*<sup>11</sup>, on which the existing subnational GDP dataset<sup>8</sup> is based. We then updated subnational data for nearly every country for which





subnational data were available, resulting in data for 2708 subnational units. Moreover, we expanded the temporal coverage of subnational units from 1990–2010 to 1990–2021. We used data from existing databases, such as OECD<sup>24</sup> and Eurostat<sup>25</sup>, DOSE v2 (ref. <sup>10</sup>), and national censuses. The source for each dataset used with the years it covers is given in the Supplementary Data File. For some countries, we combined data from several different sources (for different years); each source is given with the corresponding temporal coverage.

The number of available years with data varied greatly between the countries (Fig. 2a), as did the range of years (Fig. 2b). However, the mean interval was for most of the countries (56/89) 1, meaning that there were data for every year within the range of reported years and 2 or fewer for 72/89 countries (Fig. 2c). Notably, for the African continent, the data availability is very low, whereas in all other regions, the geographical data coverage is rather good (Fig. 2).

Due to changes in administration areas within some countries—some admin 1 level areas were split or combined within the study area (we used the www.statoids.com webpage to obtain information on these changes)—we needed to estimate the values for some administrative areas for a limited number of years. These are documented in the Supplementary Data File.

Once the subnational data were collected, we calculated the population-weighted national average GDP per capita values for each country and reported year. For population data, we used GHS-POP R2023A global gridded population data<sup>26</sup>. The data are given every five years with a 30 arc-sec scale. We aggregated it to a 5 arc-min scale and interpolated the missing years using linear interpolation. The derived Admin 0 level GDP per capita data were then used to calculate the ratio between the Admin 1 level and Admin 0 level GDP per capita values (subnational GDP ratio). By doing so, we were able to combine the different datasets regardless of whether the data were in USD or local currency. Additionally, later in the process, when we at the very end multiplied the ratio with reported Admin 0 level data, we ensured that the national total GDP was equal to the sum of that of subnational areas; i.e., this harmonised the admin 1 level data with reported Admin 0 level data.







Fig. 5 Downscaling model error of each administrative unit. Differences between reported and cross-validated predicted values for (a) first sampled year and (b) last sampled year.

.....

The subnational GDP ratios were interpolated between the years with reported values using the na.approx function in R (under the zoo package<sup>23</sup>). For extrapolation, we used the nearest observed value, i.e., assuming that the relationship between the subnational areas and the admin 0 level GDP per capita remained constant (i.e., the admin 0 level data were used to scale the admin 1 level data in these areas, as in Kummu *et al.*<sup>8</sup>). As a



**Fig. 6** Downscaling error (reported–predicted) of the subnational GDP ratio for each administrative area–year combination for 12 different meta-regions.

result, we obtained gap-filled full time series of subnational GDP ratios for all the countries with reported admin 1 level data. This was used to train the downscaling data (see next section).

**Downscaling to the admin 2 level.** We downscaled (or disaggregated) the combined subnational GDP ratio (i.e., admin 1 level/admin 0 level) dataset to admin 2 level resolution using a set of machine learning algorithms. We randomly selected 50% of the years with reported admin 1 level data from each country. This resulted in n = 25,608 observations.

*Independent variables.* We then prepared independent datasets for downscaling (gridded urbanisation level, gridded travel time to the closest city, income inequality at the admin 0 level, and GDP at the admin 0 level). All these data were aggregated to the admin 1 level (using the administrative boundary sequestration of the reported GDP data) by calculating the population-weighted average for each administrative unit.

*Urbanisation level.* The urbanisation level is mainly positively linked to GDP per capita, as shown, for example, by Chen *et al.*<sup>27</sup>. We developed our own urbanisation level dataset by using the same population count data as those used to estimate the admin 0 level GDP per capita from the admin 1 level data (see Section 2.2), i.e., GHS-POP R2023A<sup>26</sup>.

To obtain the urbanisation level, we combined the gridded population count data with the national-level urbanisation rate from the World Urbanisation Prospects (2018 revision)<sup>28</sup> as follows.

- 1. We first estimated the annual urbanisation rate from the World urbanisation prospects data (which includes data for each five-year period) and then interpolated and extrapolated the data for each year using the same method as that used for the admin 0 level GDP per capita (see above).
- 2. We then used the gridded population density and count data (see above) separately for each country and year to determine the cumulative population starting from the grid cell with the lowest population density to the highest, and we calculated the cumulative percentage of the total national population.
- 3. We then compared this to the national urbanisation rate, i.e., if the urbanisation rate was 43% for a given year, the threshold for population density for urban areas is the one that is reached with a cumulative percentage of the total population of 1-0.43 = 0.57. This population density is assigned a value of 1.
- 4. Finally, we scaled the population density data so that the grids with the highest density were assigned a



Error: reported subnational GDP ratio - predicted subnational GDP ratio

Fig. 7 Error histogram of the downscaling model (reported – predicted) of the subnational GDP ratio for each of the 12 different meta-regions.

value of 2, those with a threshold value of 1 and those with the lowest density were assigned a value of 0. Thus, all rural areas have a value < 1, and all urban areas have a value > 1.

If we could not estimate the urbanisation level for an administrative unit in question (this was the case for some small island states and sovereign territories), we used a value of 1 (on a scale of 0 to 2).

*Travel time to the closest city.* The travel time to the closest city, which is highly linked to road infrastructure, has been shown to have an impact on economic activities<sup>29</sup> and thus was selected as one of the independent variables. We used raster data of travel time to the closest city of at least 50,000 people from<sup>30</sup> with a 1 km resolution for 2015. We resampled it to 5 arc-min resolution so that we were able to weight it with the population dataset when aggregating it to the administrative units.

Income inequality at the admin 0 level. We used income inequality (Gini coefficient) as a proxy for how evenly income is distributed across the country and thus was used as an indicator of potential heterogeneity in the GDP ratio between the admin 1 level units. We used the national-level Gini coefficient, compiled by authors from multiple sources, for a total of n = 198 countries<sup>6</sup>. If a country did not have data for a Gini coefficient (this was the case for some small island states and sovereign territories), we used a value of 0.5 (on a scale from 0 to 1).

GDP per capita for the admin 0 level. This independent variable, GDP per capita (PPP) at the admin 0 level, represents the economic status of a country where the admin 1 level unit is located. The reason for adding this to the downscaling was to take into account the different dynamics of how the GDP ratio varies within a country depending on the overall economic level of the country. As we downscaled the admin 1 level/ admin 0 level ratio, these admin 0 level GDP per capita data were independent from the ratio data.

Downscaling procedure. After analysing the selected admin 1 level GDP ratio data for downscaling, we retained only the observations where the GDP ratio was less than or equal to 5 (i.e., we removed the outliers from the data).



**Fig. 8** Comparison of the admin 0 level, admin 1 level and admin 2 level GDP per capita (PPP) products. (a) GDP per capita for the admin 0 level; (b) slope for the admin 0 level; (c) GDP per capita for the admin 1 level; (d) slope for the admin 1 level; (e) GDP per capita for the admin 2 level; and (f) slope for the admin 2 level. The GDP per capita is for 2022. Slope was estimated using the Siegel repeated medians method (calculated with R's mblm package) over the study period (1990–2022). All maps are overlaid with country (admin 0 level) boundaries from Natural Earth data.



**Fig. 9** Total GDP (PPP) for 2022 in million USD. Plotted from the 5 arc-min resolution product, overlaid with country (admin 0 level) boundaries from Natural Earth data. Note: log2 scale for the colour scale.

SCIENTIFIC DATA (2025) 12:178 | https://doi.org/10.1038/s41597-025-04487-x



**Fig. 10** Validation of the downscaled admin 2 level data. Reported admin 2 level GDP per capita (PPP) for Belgium and Croatia compared against downscaled GDP per capita (PPP) for the same administrative areas. (a) all data; (b) the highest values were removed to test how they impact the correlation result. Correlation analysis was conducted with the Pearson correlation. The number of individual admin 2 level units is 43.



**Fig. 11** Validation of the aggregated admin 1 level (interpolated and extrapolated) and admin 2 level (downscaled) data against reported national level data. (**a**) aggregated admin 1 level data compared with reported national data; (**b**) aggregated admin 2 level data compared with reported national data. Correlation analysis was conducted with the Pearson correlation. Aggregation from subnational to national level was performed using the weighted population mean. The number of individual admin 0 level units is 235.

.....

A rigorous search was conducted among some of the well-established machine learning algorithms to find the most suitable model for our purposes. After a preliminary analysis, multilayer perceptron (MLP)<sup>31</sup>, support vector regression (SVR)<sup>32</sup>, and ensembled trees<sup>33</sup> were chosen for further study. The dataset (n = 25,307) consisted of coded data from 12 different meta-regions, following UN classification similar to that of Kummu *et al.*<sup>34</sup>. Eighty percent (80%) of the data from each of the 12 regions were randomly sampled to create the training set (n = 20,245). The remaining 20% of the data from each region were then used to create the testing set (n = 5,062), maintaining the balanced representation of regions in both sets. Each model has a set of hyperparameters to tune. To find the best hyperparameter sets for each model, we ran Bayesian optimisation<sup>35</sup>. We chose this method over grid search to avoid testing every possible configuration and tested only those that were more likely to improve the performance of the models. The ensembled trees model with 152 learners using the least-squares boosting method with a learning rate of 0.07 was chosen because it outperformed all the other models in both the training and testing phases (Fig. 3). Although we can see that the training accuracy ( $R^2 = 0.80$ ) for this model is significantly greater than the tof the runner-up MLP model ( $R^2 = 0.49$ ), it is still the superior model.

*Downscaling error.* To assess the downscaling error, we first mapped the subnational error (predicted subnational GDP ratio – reported subnational GDP ratio) for each admin unit (Fig. 5). We chose the first and last years



**Fig. 12** GDP per capita (PPP) for three different geographical areas. (**a,d,g**) South–Southeast Asia; (**b,e,h**) Central South America; and (**c,f,i**) Central–Southern Europe. (**a**–**c**) admin 0 level; (**d**–**f**) admin 1 level; and (**g**–**i**) admin 2 level. Plotted from the polygon GDP per capita (PPP) product, it is overlaid with country (admin 0 level) boundaries from Natural Earth data. Note: log10 scale for the colour scale.

.....

sampled for training the downscaling model (years differ from country to country, see above), and it seems that the error is smaller the closer we are to the end of the study period. In general, the error is smallest in Europe and North America and largest in Central Asia and Latin America.

Furthermore, to assess the error in the downscaling method and how it differs from region to region, we used the 12 meta-regions (see the Downscaling procedure section) and plotted the error for each year-admin area combination (Fig. 6). These results show that, in general, the downscaling model predicts the subnational GDP ratio rather well, with the RMSE ranging from less than 0.15 (Australia & Oceania, North America and Europe) to 0.25-0.3 (Eastern Europe and Central Asia and Southeast Asia). Finally, we plotted the histogram for errors for each meta-region, as presented in Fig. 7.

**Prediction for the admin 2 level.** To predict the admin 2 level GDP ratio (admin 2 level vs admin 1 level), we used the trained model with the independent variables aggregated to the admin 2 level. We obtained the administrative boundary layer for the admin 2 level from the GADM database (https://gadm.org). If there was no admin 2 level division for a country, we used the admin 1 level division instead. If this information was not available, then we used the admin 0 level division.



**Fig. 13** Total GDP (PPP) for 2020 in million USD for three geographical areas. (**a**) South–Southeast Asia, (**b**) Central South America, (**c**) Central–Southern Europe. Plotted from the 5 arc-min resolution product, overlaid with country (admin 0 level) boundaries from Natural Earth data. Note: log2 scale for the colour scale.

.....

The data for the independent variables were the same as those used for training the downscaling model, with the exception that instead of using the admin 0 level data for the Gini coefficient<sup>6</sup> and GDP per capita, we used here the admin 1 level data (i.e., again one administrative level higher than the administrative area in question). The validation for downscaled admin 2 level GDP per capita data is given in Section 4.

**Data harmonisation.** To ensure consistency and comparability between subnational and national economic data (admin 1 and downscaled admin 2 levels), we implemented a data harmonisation procedure. The GDP per capita in purchasing power parity (PPP) was utilised to adjust for price level differences and variations in the cost of living between countries.

Subnational GDP per capita PPP data were collected from diverse sources and currencies, which we harmonised by calculating ratios relative to the national average (derived as the population-weighted average of the subnational data). This ratio-based approach, rather than using absolute values, was employed for interpolation, extrapolation, and downscaling. After downscaling, absolute subnational GDP per capita PPP values were computed based on the reported national figures.

To ensure further full alignment with national data, at the very end, we calculated the ratio between the population-weighted national average derived from subnational data (admin 1 level or admin 2 level) and the reported national data. This ratio was then used to correct biases in the subnational data. This detailed procedure, as depicted in Fig. 1, ensures comparable economic data across various regions that are harmonised with national reported data.

**Total GDP (PPP) data.** We used GDP per capita Admin 2 level data to estimate total GDP (PPP) by multiplying the per capita data by the population count. We did this for three resolutions: 30 arc-sec (ca. 1 km at the equator), 5 arc-min, and 30 arc-min. To produce the 30 arc-sec product, we resampled the 5 arc-min GDP per capita data to that resolution.

For the 30 arc-sec product, we used GHS population grid (GHS-POP R2023A) data<sup>26</sup> for each year for which population data were available (1990, 1995..., 2015, 2020). For the 5 arc-min product, we used the same GHS-POP R2023A population dataset (as with other parts of the dataset creation), aggregated to 5 arc-min resolution. Finally, we resampled the 5 arc-min data to 30 arc-min resolution.

#### **Data Records**

The data are available at the following online repository: https://doi.org/10.5281/zenodo.10976733<sup>36</sup>. We provide the following data:

- GDP per capita (PPP) at the admin 0 level (national) for 1990-2022 (GeoTIFF, gpkg, csv)
- GDP per capita (PPP) at the admin 1 level (at the level of reporting, either the administration 1 level or the admin 0 level) for 1990–2022 (GeoTIFF, gpkg, csv)
- GDP per capita (PPP) at the admin 2 level (downscaled from the admin 1 level) for 1990–2022 (GeoTIFF, gpkg, csv)
- Total GDP (PPP), downscaled admin 2 level GDP per capita (PPP) multiplied by the gridded population count for 1990–2022, with three resolutions: 30 arc-sec, 5 arc-min, and 30 arc-min (GeoTIFF)
- The input data for the script were used to generate the data above (code\_input\_data.zip). The code is available at https://github.com/mattikummu/griddedGDPpc

The metadata with reported data years and sources for both the admin 0 and admin 1 levels are given in the Supplementary Material of this article.

The GDP per capita (PPP) data product for three administrative levels for 2022 and the slope over the study period of 1990–2022 using the Siegel repeated medians method (calculated with R's mblm package) are shown in Fig. 8. The gridded total GDP (PPP) for 2022 is shown in Fig. 9.

#### **Technical Validation**

We validated the final product at the admin2 level (downscaled from the admin 1 level) using subnational GDP per capita (PPP) data from OECD regional statistical databases (ref). The data include admin 2 level data for Belgium and Croatia (n = 43 subnational units), which were not used to create the dataset. These observed OECD data were compared against the aggregated downscaled GDP per capita (PPP) for the same administrative units.

After aggregation, we harmonised the OECD subnational data with the national GDP per capita (PPP) data of our dataset and the OECD dataset. Then, we sampled 25% of the observed years and compared the down-scaled values to the observed values using Pearson correlation (Fig. 10). The results show high accuracy, with a Pearson R of 0.88 (and without the highest GDP per capita points of 0.84).

Furthermore, we validated how well our harmonisation of the subnational data products against reported national data by the World Bank, IMF and CIA worked. We did this by first aggregating the admin 1 level and admin 2 level data to the national level (admin 0 level) using population weighted average mean. We then compared these data with reported national-level data. We found that harmonisation had worked as intended, resulting in a Pearson's R of 1 (p < 0.001) for both admin levels (Fig. 11).

#### **Usage Notes**

To show the usability of the data, we plotted the admin 0 level, admin 1 level and admin 2 level GDP (PPP) per capita data for three selected geographical areas (Fig. 12) for 2022. This demonstrates how much variability there is within shown countries and even within the 1-level administrative units. Furthermore, we also mapped the total GDP (PPP) for 2022 for the three selected geographical areas (Fig. 13), demonstrating better details than the results shown in a global map (Fig. 9).

#### **Code availability**

The code is available at the following github repository: https://github.com/mattikummu/griddedGDPpc. The input data for the code are available in the data repository: https://doi.org/10.5281/zenodo.10976733.

Received: 25 April 2024; Accepted: 15 January 2025; Published online: 30 January 2025

#### References

- 1. Mahecha, M. D. et al. Earth system data cubes unravel global multivariate dynamics. Earth System Dynamics 11, 201-234 (2020).
- 2. Ward, P. J. et al. Review article: Natural hazard risk assessments at the global scale. Natural Hazards and Earth System Sciences 20, 1069–1096 (2020).
- 3. Otto, I. M. et al. Socio-economic data for global environmental change research. Nature Clim Change 5, 503-506 (2015).
- Niva, V. et al. World's human migration patterns in 2000–2019 unveiled by high-resolution data. Nat Hum Behav 1–15 https://doi. org/10.1038/s41562-023-01689-4 (2023).
- 5. Liu, X. et al. High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. Nat Sustain 3, 564–570 (2020).
- Chrisendo, D. et al. Income inequality has increased for over two-thirds of the global population. PREPRINT (Version 1) available at Research Square https://doi.org/10.21203/rs.3.rs-5548291/v1 (2024).
- 7. Smits, J. & Permanyer, I. The Subnational Human Development Database. Sci Data 6, 190038 (2019).
- Kummu, M., Guillaume, J. H. A. & Taka, M. Gridded global datasets for Gross Domestic Product (GDP) and Human Development Index (HDI) over 1990-2015. Scientific Data 5, 180004 (2018).
- 9. Wang, T. & Sun, F. Global gridded GDP data set consistent with the shared socioeconomic pathways. *Sci Data* 9, 221 (2022).
- 10. Wenz, L., Carr, R. D., Kögel, N., Kotz, M. & Kalkuhl, M. DOSE Global data set of reported sub-national economic output. Sci Data 10, 425 (2023).
- Gennaioli, N., La Porta, R., Lopez-de-Silanes, F. & Shleifer, A. Human Capital and Regional Development. *The Quarterly Journal of Economics* 128, 105–164 (2013).
- 12. Chrisendo, D., Piipponen, J., Heino, M. & Kummu, M. Socioeconomic factors of global food loss. Agriculture & Food Security 12, 23 (2023).
- Rizzati, M. et al. The local costs of global climate change: spatial GDP downscaling under different climate scenarios. Spatial Economic Analysis 18, 23–43 (2023).
- 14. Herrera-García, G. et al. Mapping the global threat of land subsidence. Science 371, 34-36 (2021).
- Jongman, B., Ward, P. J. & Aerts, J. C. J. H. Global exposure to river and coastal flooding long term trends and changes. *Global Environmental Change* 22, 823–835 (2012).
- Zhong, C. et al. Evaluating trends, profits, and risks of global cities in recent urban expansion for advancing sustainable development. Habitat International 138, 102869 (2023).

- Farooq, H., Azevedo, J. A. R., Soares, A., Antonelli, A. & Faurby, S. Mapping Africa's Biodiversity: More of the Same Is Just Not Good Enough. Systematic Biology 70, 623–633 (2021).
- Dasgupta, S., Emmerling, J. & Shayegh, S. Inequality and growth impacts of climate change—insights from South Africa\*. Environ. Res. Lett. 18, 124005 (2023).
- 19. Tian, Y., Tsendbazar, N.-E., van Leeuwen, E. & Herold, M. Mapping urban-rural differences in the worldwide achievement of sustainable development goals: land-energy-air nexus. *Environ. Res. Lett.* **17**, 114012 (2022).
- 20. World Bank. World Development Indicators (2023).
- 21. IMF. World Economic Outlook GDP per Capita PPP. https://www.imf.org/external/datamapper/PPPPC@WEO (2023).
- 22. IndexMundi. IndexMundi Country Facts. https://www.indexmundi.com/ (2023).
- 23. Zeileis, A. & Grothendieck, G. 200: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software* 14, 1–27 (2005).
- 24. OECD. OECD Regional Database. https://stats.oecd.org/Index.aspx?DataSetCode=REGION\_DEMOGR (2024).
- 25. Eurostat. Database Eurostat. https://ec.europa.eu/eurostat/data/database (2023).
- 26. European Commission. GHSL Data Package 2023. https://data.europa.eu/doi/10.2760/098587 (2023).
- Chen, M., Zhang, H., Liu, W. & Zhang, W. The Global Pattern of Urbanization and Economic Growth: Evidence from the Last Three Decades. PLOS ONE 9, e103799 (2014).
- 28. United Nations. World Urbanisation Prospects: The 2018 Revision. (2019).
- 29. Meijer, J. R., Huijbregts, M. A. J., Schotten, K. C. G. J. & Schipper, A. M. Global patterns of current and future road infrastructure. *Environ. Res. Lett.* 13, 064006 (2018).
- 30. Nelson, A. et al. A suite of global accessibility indicators. Sci Data 6, 266 (2019).
- Ahmed, K., Shahid, S., Haroon, S. B. & Xiao-jun, W. Multilayer perceptron neural network for downscaling rainfall in arid region: A case study of Baluchistan, Pakistan. J Earth Syst Sci 124, 1325–1341 (2015).
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. & Vapnik, V. Support Vector Regression Machines. in Advances in Neural Information Processing Systems vol. 9 (MIT Press, 1996).
- 33. Breiman, L. Bagging predictors. Mach Learn 24, 123–140 (1996).
- Kummu, M. et al. The world's road to water scarcity: shortage and stress in the 20th century and pathways towards sustainability. Scientific Reports 6, 38495 (2016).
- Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. in Advances in Neural Information Processing Systems (eds. Pereira, F., Burges, C. J., Bottou, L. & Weinberger, K. Q.) vol. 25 (Curran Associates, Inc., 2012).
- Kummu, M. et al. Data for: Downscaled gridded global dataset for Gross Domestic Product (GDP) per capita at purchasing power parity (PPP) over 1990–2022, Zenodo, https://doi.org/10.5281/zenodo.10976733 (2024).

#### Acknowledgements

This study was funded by the European Research Council under the European Union's Horizon 2020 research and innovation program grant 819202 (M. Kummu), the Research Council of Finland's project TREFORM 339834 (M. Kummu), the Research Council of Finland's Flagship Programme under project Digital Waters 359248 (M. Kummu), Maa- ja vesitekniikan tuki ry (S. Masoumzadeh Sayyar), and The World Bank (financial support for the data work). The authors would like to thank Dr Daniel Chrisendo, Dr Venla Niva and Dr Vilma Sandström for commenting on the manuscript. ChatGPT v4 and Curie were used to proofread the manuscript.

#### **Author contributions**

M. Kummu and M. Kosonen collected the reported subnational data. M. Kummu developed the code to create the dataset with some support from M. Kosonen. S. Masoumzadeh Sayyar performed the downscaling part of the study. M. Kummu performed the technical validation. M. Kummu and S. Masoumzadeh Sayyar provided the illustrations. M. Kummu wrote the manuscript with support from S. Masoumzadeh Sayyar.

#### **Competing interests**

The authors declare no competing interests.

#### **Additional information**

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04487-x.

Correspondence and requests for materials should be addressed to M.K.

Reprints and permissions information is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025