nature sustainability

Supplementary information https://doi.org/10.1038/s41893-025-01689-4

Rising income inequality across half of global population and socioecological implications

In the format provided by the authors and unedited

Supplementary notes

Supplementary for Chrisendo et al. (Rising income inequality across half of global population and socioecological implications)

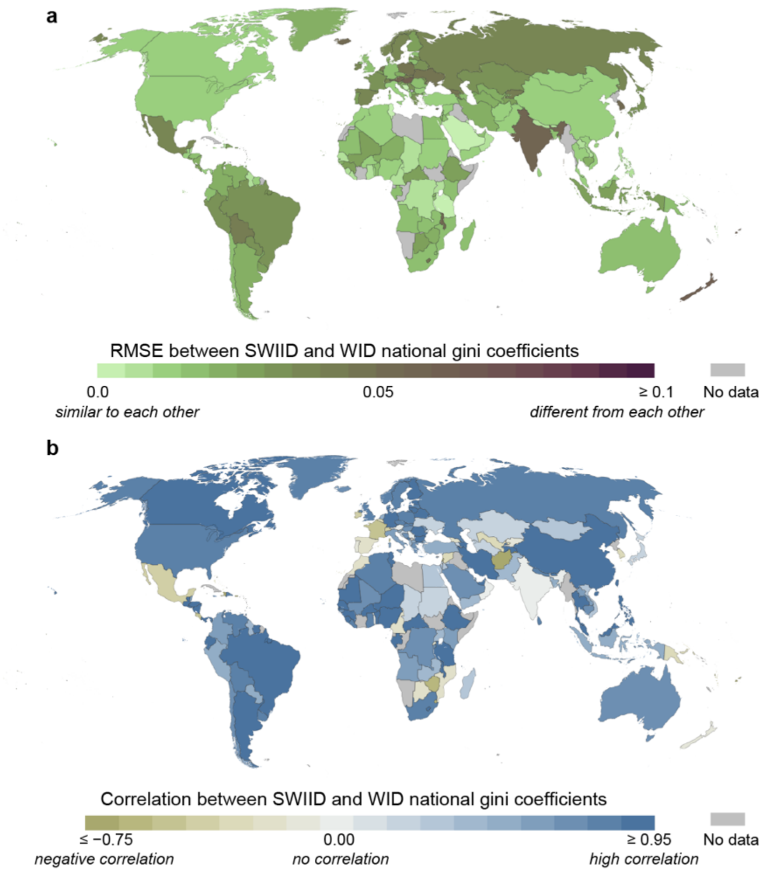
1 National level data comparison

National-level inequality data are available from three main sources: the World Bank (WB), the World Inequality Database (WID), and the Standardized World Income Inequality Database (SWIID) (Methods of the main text). Of these, the WB dataset has much less reported data than the two other datasets and therefore, we focus here on the WID and SWIID datasets.

To estimate the differences between these datasets, we compared the temporal patterns of Gini estimates between SWIID and WID. For the comparison, we selected countries with at least 10 overlapping observations. A direct comparison is not possible, as the methods used to generate these products differ (Methods of the main text). Thus, for each country, we first normalised the Gini time series by dividing each entry by the national average over time. Then, we computed the correlation and Root Mean Square Error (RMSE).

We found that in some countries, these two datasets agree very well (i.e., the RMSE of the normalised value is small and the correlation is high); these countries include, for example, China, the USA, Germany, and many countries in Southeast Asia, as well as Africa (Supplementary Figure 1). For most countries, these two datasets agree relatively well, with a low RMSE and a high correlation. However, for a few countries, such as India, Mexico, and Ukraine, the RMSE is high, and the correlation is negative. For some countries, different signs are observed; for example, in Mongolia, the RMSE is low, but the correlation is also low.

The differences are expected, as the two datasets are using different data as input for the Gini coefficient. SWIID uses a wide array of published Gini coefficients, and the Luxembourg Income Study (LIS) serves as the gold standard for ensuring comparability. WID, in turn, relies on microdata from surveys, administrative (tax) data, and national accounts data. Therefore, we decided to derive the subnational Gini coefficient data using both datasets.



Supplementary Figure 1. Agreement of the SWIID and WID datasets. These two national-level Gini coefficient datasets were compared for countries with at least 10 overlapping observations. a) harmonic score between the datasets, a combination of root mean square error and correlation; b) correlation between the datasets.

2 Uncertainty in Gini coefficient data

To capture the uncertainty in the Gini coefficient and in the derived slope over the study period (1990-2023), we employed a Monte Carlo simulation. This was applied to both national (admin 0) and subnational (admin 1) levels, accounting for various sources of uncertainty and data sparsity. The analysis was conducted using R statistical software (version 4.3.2, R Core Team, 2023).

2.1 Data

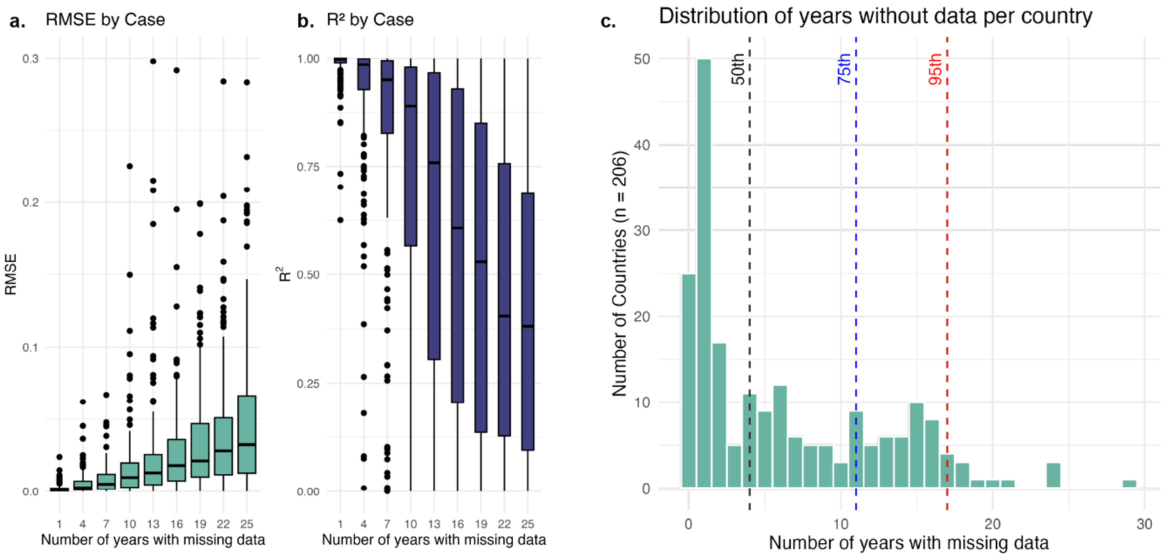
For the admin 0 level, we used data from SWIID, as that includes the standard error for each data entry[1]. For admin 1 level, we used the data compiled in this study.

2.2 Uncertainty originating from extrapolation

We started the uncertainty analysis by assessing the uncertainty originating from national-level extrapolation (see Methods). To quantify the uncertainty, we conducted a sensitivity analysis using a leave-out strategy applied to the nearly complete dataset of Gini coefficients. We systematically simulated missing data scenarios by removing varying numbers of initial years (1 to 25 in steps of 3) from the time series of each country, which had complete or nearly complete data coverage over the study period. For each scenario, the first n years were set to NA for a focal country, and its remaining data were excluded from the reference dataset to ensure unbiased extrapolation.

Using the extrapolation function (f_extrapol.R) adopted from Kummu et al (2025)[2], we predicted missing values based on the remaining dataset (see Methods). The performance of extrapolated values was evaluated for each country and scenario using RMSE and Coefficient of Determination (R^2).

The results show that the RMSE increases with the number of years of missing data, with the median reaching approximately 0.01 with 10 years of missing data and 0.03 with 25 years of missing data (Supplementary Figure 2a). The range also increases gradually. The R2, in turn, decreases with the number of years with missing data, remaining relatively high (over 0.75) for less than 13 years with missing data, but only approximately 0.35 for 25 years of missing data (Supplementary Figure 2b). However, when looking at the data coverage (Supplementary Figure 2c), 95% of the countries have less than 17 years of missing data.



Supplementary Figure 2. Uncertainty of the extrapolation. Results from the sensitivity analysis using a leave-out strategy in which we systematically simulated missing data scenarios by removing varying numbers of initial years (1 to 25 in steps of 3) from the time series. a) root mean square error (RMSE) over the countries in the uncertainty analysis (complete or nearly complete data coverage over time), for each assessed number of years with missing data; b) R2 over the countries in the uncertainty analysis (complete or nearly complete data coverage over time), for each assessed number of years with missing data; c) histogram of the SWIID dataset, showing the number of years with missing data in the reported Gini coefficient data. In the boxplots, the whiskers represent the 5th-95th percentile, and the boxes represent the 25th-75th percentile.

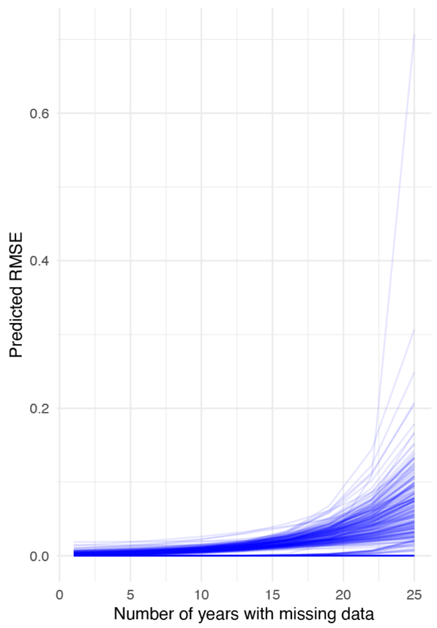
2.3 Monte Carlo simulation framework

A Monte Carlo simulation with n = 500 runs was performed for each country. For countries with subnational data, both admin 0 and admin 1 levels were simulated concurrently within each run to ensure consistency. To generate plausible alternative data series, we introduced simulated noise, reflecting both measurement error and inherent variability, and then we calculated Gini statistics and trends for each simulated series.

2.3.1 Standard error estimate

To assess the standard error (SE), we used the reported SE for admin 0 areas with data. For each Monte Carlo run, we randomly sampled a specific RMSE curve from a pre-defined rmse_matrix, which represents various potential RMSE-distance relationships (see Supplementary Figure 3). This curve was then used to quantify the expected increase in uncertainty with temporal distance from a reported observation.

Admin 0 Level: If a Gini value for a year in question was reported, the reported SE was used. If there was no data, we calculated the adjusted_gini_disp_se by adding the sampled RMSE value—corresponding to the temporal distance from the nearest reported Gini observation (see curves in Supplementary Figure 3)—to the SE of the nearest reported observation. This method accounted for the uncertainty in the extrapolation.



Supplementary Figure 3. Ensemble of Log-Linear RMSE curves.

Admin 1 Level: Error estimates are not available for the admin 1 level data. Therefore, for years with reported admin 1 data, we used the same SE as the admin 0 level data. For the missing years, we did not extrapolate the admin 1 level data. Instead, we used the ratio (admin 1 / admin 0) from the closest admin 1 level data. For the years with missing data, we added an estimation error using the admin 0 extrapolation error to give a range. We sampled from a truncated normal distribution that is based on the admin 0 extrapolation error, as there is no reference data for admin 1 level data that could be used to derive the error range (mean = 0.002, SD = 0.0005, constrained between 0.001 and 0.003) to represent an irreducible, additional standard error inherent to subnational estimates.

If an admin 1 Gini was directly reported, its base SE came from the admin 0 data for the same year. For missing admin 1 values, we used the nearest reported admin 1 observation (derived from admin 0's SE) as a base. We then scaled an additional component with the distance to the reported admin 1 data entry. This method ensures all admin 1 estimates have a base of additional uncertainty, and imputed values gain more uncertainty the further they are from reported data.

The total admin 1 standard error (σ_{adm1}) for the simulation was calculated by combining the variance from the ratio and the adm0 standard error using the error propagation formula:

$$\sigma_{adm1} = \sqrt{(r \times \sigma_{adm0})^2 + (G_{adm0} \times \sigma_{ratio})^2} \quad (1)$$

Where σ_{ratio} is derived from the ratio between admin 1 and admin 0 Gini coefficient as follows:

$$\sigma_{ratio} = r \times \sqrt{\left(\frac{\sigma_{adm1}}{G_{adm1}}\right)^2 + \left(\frac{\sigma_{adm0}}{G_{adm0}}\right)^2} \quad (2)$$

In equations (1) and (2), the variables G_{adm0} and G_{adm1} represent the Gini dispersion estimates at the admin 0 and 1 levels. The terms σ_{adm0} and σ_{adm1} denote the standard errors of G_{adm0} and G_{adm1}, respectively. r refers to the ratio between the admin1 and admin0 Gini estimates. In error propagation, we assume that the two error sources, namely the admin 0 and 1 level SEs, are statistically independent; therefore, the covariance term in the general error-propagation formula is omitted.

2.3.2 Gini coefficient simulation

For each simulation run, we generated time series of Gini coefficients by adding normally distributed, autocorrelated noise to the base Gini values. To account for autocorrelation between years, we used a coefficient of $\rho=0.9$. The error vector for each time series was generated using a multivariate normal distribution (mvrnorm from the MASS package), with a mean of zero and a covariance matrix derived from ρ and the calculated standard errors for each year. This simulated error was then added to the reported Gini values (or ratio-adjusted values for admin 1). This process created a new, plausible time series for each Gini measure.

2.3.3 Slope estimation

For each simulated Gini time series (both admin 0 and admin 1, per simulation run), we calculated Siegel repeated medians slope using the mblm package. Siegel's slope is a robust, non-parametric method for estimating trend slopes. It is less sensitive to outliers than ordinary least squares regression. We set a non-significant slope ($p > 0.1$) to zero to avoid interpreting non-significant trends.

2.3.4 Output and aggregation

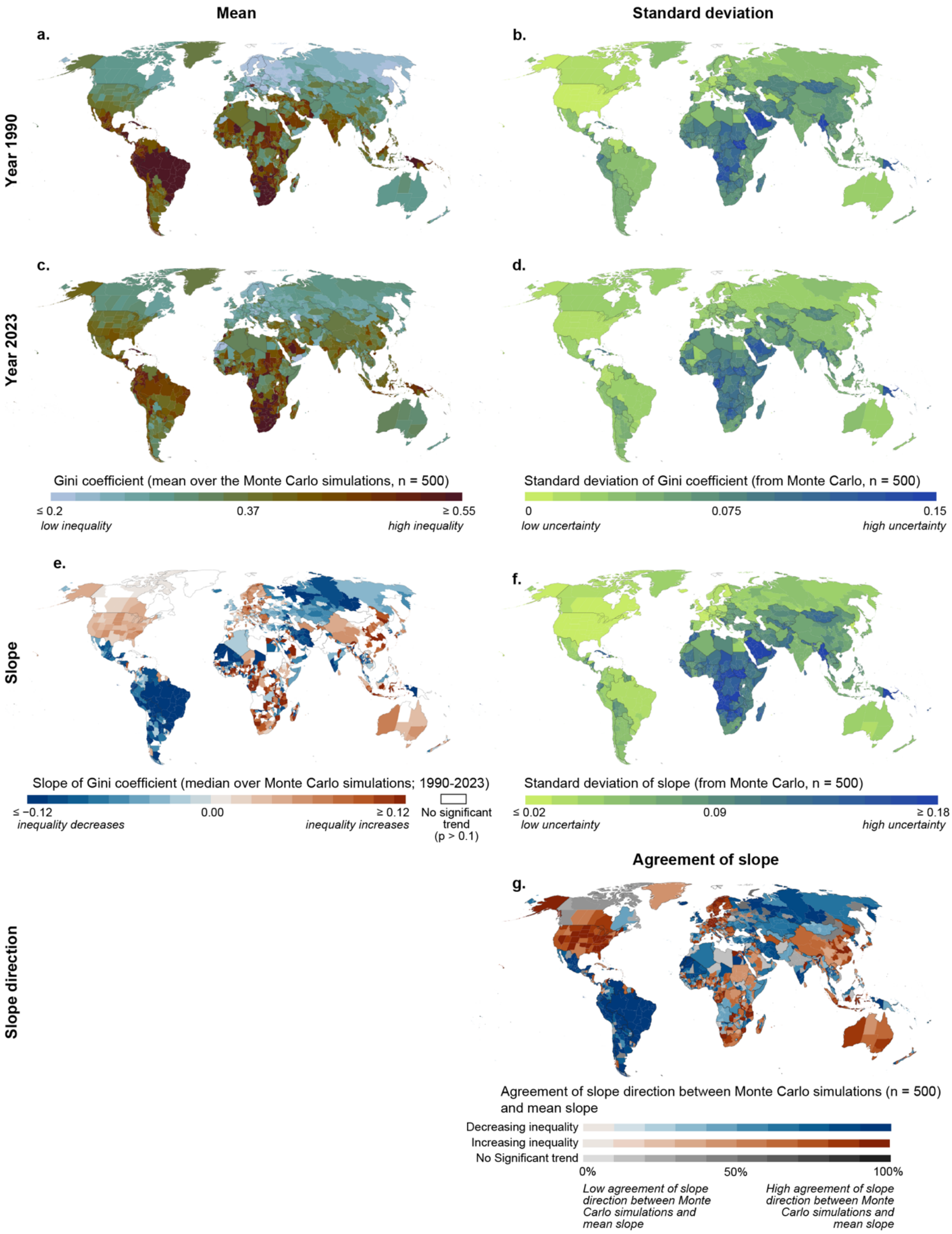
For each admin area, the Monte Carlo simulation produced:

- Mean and standard deviation of the simulated Gini coefficients across all runs, per year
- Individual Siegel's slopes for each simulation run. These allow assessment of the distribution of potential trends.

2.4 Uncertainty estimates

We found, not surprisingly, that uncertainty for the Gini coefficient is high where data are scarce and often of poor quality. This is the case in Africa and the Middle East (Supplementary Figure 4b & d). The uncertainty is lowest in North America and Europe. In most regions, uncertainty is higher for 1990 compared to 2023, as data coverage is often better the closer we are to the present date.

For slope, uncertainty is also highest in Africa and in parts of Asia (Supplementary Figure 4f). When examining agreement in slope directions (negative, zero, positive) across the Monte Carlo runs compared to the mean slope, we identified areas with nearly full agreement, such as Latin America and parts of Europe (Supplementary Figure 4g). In some countries, such as the USA, China, and India, there is high heterogeneity in agreement on slope direction. In general, areas with no significant slope exhibit poor agreement, as the Monte Carlo simulations reveal an almost equal distribution across all slope directions (negative, positive, or no significant trend).



Supplementary Figure 4. Uncertainty results of Monte Carlo simulations. a) mean of Gini coefficient for year 1990; b) standard deviation of Gini coefficient for year 1990; c) mean of Gini coefficient for year 2023; d) standard deviation of Gini coefficient for year 2023; e) mean slope of Gini coefficient over 1990-2023; f) standard deviation for slope of Gini coefficient over 1990-2023; g) agreement between mean slope and all runs of the Monte Carlo simulation (n = 500). Slopes were calculated with Siegel repeated medians (mblm R package), and only statistically significant slopes (p < 0.1) are reported.

References

1 Solt F. Measuring income inequality across countries and over time: The standardized world income inequality database. *Soc Sci Q* 2020; **101**: 1183–1199.

2 Kummu M, Kosonen M, Masoumzadeh Sayyar S. Downscaled gridded global dataset for gross domestic product (GDP) per capita PPP over 1990–2022. *Sci Data* 2025; **12**: 178.