

LETTER • OPEN ACCESS

## Non-parametric projections of national income distribution consistent with the Shared Socioeconomic Pathways

To cite this article: Kanishka B Narayan *et al* 2023 *Environ. Res. Lett.* **18** 044013

View the [article online](#) for updates and enhancements.

You may also like

- [YZICS: Unveiling the Quenching History of Cluster Galaxies Using Phase-space Analysis](#)

Jinsu Rhee, Rory Smith, Hoseung Choi et al.

- [Forest fire effects on stream water quality at continental scales: a meta-analysis](#)

Tyler B Hampton, Simon Lin and Nandita B Basu

- [The energy and carbon inequality corridor for a 1.5 °C compatible and just Europe](#)

Ingram S Jaccard, Peter-Paul Pichler, Johannes Többen et al.

ENVIRONMENTAL RESEARCH  
LETTERS

## LETTER

## OPEN ACCESS

RECEIVED  
22 August 2022REVISED  
6 February 2023ACCEPTED FOR PUBLICATION  
21 February 2023PUBLISHED  
21 March 2023

Original content from  
this work may be used  
under the terms of the  
[Creative Commons  
Attribution 4.0 licence](#).

Any further distribution  
of this work must  
maintain attribution to  
the author(s) and the title  
of the work, journal  
citation and DOI.

Non-parametric projections of national income distribution  
consistent with the Shared Socioeconomic PathwaysKanishka B Narayan<sup>\*</sup> , Brian C O'Neill , Stephanie T Waldhoff and Claudia Tebaldi 

Joint Global Change Research Institute (JGCRI), Pacific Northwest National Lab (PNNL), Washington, D.C., United States of America

<sup>\*</sup> Author to whom any correspondence should be addressed.E-mail: [kanishka.narayan@pnnl.gov](mailto:kanishka.narayan@pnnl.gov)**Keywords:** SSPs, income distributions, income inequalitySupplementary material for this article is available [online](#)

## Abstract

Understanding and projecting income distributions within countries and regions is important to understanding consumption trends and the distributional consequences of climate impacts and responses. Several global, country-level projections of income distribution are available but most project only the Gini coefficient (a summary statistic of the distribution) or utilize the Gini along with the assumption of a lognormal distribution. We test the lognormal assumption and find that it typically underestimates income in the highest deciles and over-estimates it in others. We find that a new model based on two principal components of national time series data for income distribution provides a better fit to the data for all deciles, especially for the highest and lowest. We also construct a projection model in which the first principal component is driven by the Gini coefficient and the second captures deviations from this relationship. We use the model to project income distribution by decile for all countries for the five shared socioeconomic pathways. We find that inequality is consistently higher than projections based on the Gini and the lognormal functional form, with some countries reaching ratios of the highest to lowest income deciles that are almost three times their value using the lognormal assumption.

## 1. Introduction

Projections of income distribution are of growing importance for a variety of purposes, prominently including their relevance to climate change and other environmental issues. Consumption, which drives energy use, land use, greenhouse gas emissions, and production of other pollutants, is highly dependent on the distribution of income, since demand for most commodities is non-linear in income. In addition, income distribution is a key determinant of vulnerability to environmental stressors, including to the impacts of climate change. Thus, the consequences of environmental change for society will be determined in part by future income distribution. This linkage also runs in the opposite direction: climate and other environmental stressors, as well as many environmental policies, are likely to have a greater impact on lower-income, higher-vulnerability households, exacerbating inequalities.

For example, in the climate literature, long-term projections of within-country income distribution

have been used to inform analyses of how the impacts of climate change may affect inequality and poverty (Hertel *et al* 2010, Hallegatte and Rozenberg 2017, Jafino *et al* 2020) and how inequality in turn may impact climate outcomes (Rao and Min 2018, Chen *et al* 2020). Such projections have also been applied to inform analyses of how greenhouse gas emissions mitigation may impact poverty or inequality through, for example, higher energy prices or effects on aggregate economic growth (Campagnolo and Davide 2019, Liu *et al* 2019, Fujimori *et al* 2020, Soergel *et al* 2021, Sampedro *et al* 2022).

Income distribution is also important for other, non-environmental reasons; it is a key determinant of vulnerability to social and economic stressors, and there is also an intrinsic interest in poverty and inequality outcomes that drives substantial research. The economics literature focuses primarily on understanding determinants and trends in inequality (Bourguignon 2015) and the relationship between inequality, poverty, and income growth (Ravallion 2001, Fosu 2017).

Projections of income distribution have become prominent in climate change research mainly through their association with a widely used scenario framework, the shared socioeconomic pathways (SSPs; O'Neill *et al* 2014). The SSPs are a set of five global socio-economic scenarios consisting of both qualitative (O'Neill *et al* 2017) and quantitative components that envision alternative pathways of development over the coming decades (Riahi *et al* 2017). The SSPs vary in terms of the challenges they present to adaptation to and mitigation of climate change, which translate to different combinations of demographic, economic, technological, social, and political assumptions. Inequality features prominently in these pathways; it is assumed to be particularly high in SSP3 (the 'Regional Rivalry' scenario) and SSP4 (actually named 'Inequality'), and particularly low in SSP1 ('Sustainability') and SSP5 ('Fossil-fueled Development').

The inequality aspect of the SSP narratives has been quantified by producing projections of the Gini coefficient at the national level (Rao *et al* 2019). The Gini coefficient is a summary measure of income distribution that ranges from zero (perfect equality) to one (perfect inequality). Projections of income distributions have also been produced, frequently based on the Gini coefficient scenarios. Some analyses only project specific metrics of the distribution, such as the poverty rate or Palma ratio (Hussein *et al* 2013, Campagnolo and Davide 2019). Our interest however is on projections of full distributions for a comprehensive set of deciles.

Available analyses take three broad types of approaches: (1) drawing on the existing projection of national Gini coefficients and assuming a particular functional form (usually log-normal) to describe the full distribution; (2) using a microsimulation approach in which individual household outcomes are modeled and aggregated to form income distributions (Hallegatte and Rozenberg 2017, Jafino *et al* 2020); or (3) using a combination of the first two approaches (Hughes 2019, Fujimori *et al* 2020). In addition, simpler versions of approach (1) have been taken that shift initial income distributions over time (Van der Mensbrugghe 2015, Crespo Cuaresma *et al* 2018). We provide further details on approaches (1), (2) and (3) in the supplementary information. Approach (1) has been the dominant method in the literature, so we focus on that here.

Projected Gini coefficients in many cases (e.g. Soergel *et al*, Fujimori *et al*, Hughes) have been taken from Rao *et al* (2019), although projections from Hughes (2019) and van der Mensbrugghe (2015) are also available. Rao *et al* (2019) estimate a regression model in which Gini coefficients (based on data using a mix of income concepts across countries) are driven by educational attainment (at the primary, secondary and tertiary levels), total factor productivity (TFP), and spending on health and education to produce

five alternative scenarios designed to be consistent with the SSPs. These projections show a wide range of potential outcomes that are consistent with the SSP storylines, in particular, SSPs 3 and 4 are associated with high inequality (in SSP4, especially within countries) and SSPs 1 and 5 are associated with relatively low inequality.

The log-normal distribution is the most common assumption for the functional form of the income distribution, regardless of the source of the Gini projection (Van der Mensbrugghe 2015, Riahi *et al* 2017, Hughes 2019, Fujimori *et al* 2020, Soergel *et al* 2021). It can be conveniently parameterized with the mean per capita income and Gini coefficient for a given country and thus easily used in conjunction with projections of income growth (from, for example, integrated assessment models) and Gini coefficients (SI 2 of the supplementary information provides details on the parameterization for the lognormal functional form).

However, the lognormal functional form has documented limitations. Observations are known to deviate from the lognormal in the tails of the distribution (Chotikapanich 2008, Badel *et al* 2020). At least one study of the impact of climate policy on poverty did not use the lognormal functional form when deriving poverty rates specifically for this reason (Soergel *et al* 2021). Moreover, since the Gini coefficient is more representative of the middle portion of the income distribution (Osberg 2017), specifying a lognormal distribution based on a Gini coefficient introduces further error in representing the data. In addition, two regions with the same Gini coefficient can have different underlying distributions (Chitiga *et al* 2015, Osberg 2017). Finally, studies that have used the Gini in combination with the lognormal functional form have rarely evaluated the fit of the method through comparison with the most recent data on income distributions at the country level.

Alternatives have been proposed to address the limitations of the lognormal distribution. These include different distributions such as the Weibull distribution (Hallinan Jr 1993, Rinne 2008) or hybrid parameterizations which use the lognormal for low- and middle-income levels and a pareto for high income levels (Mandelbrot 1960, Figueira *et al* 2011, Arnold 2014). In such studies that have used or proposed alternatives to the lognormal distributions, these distributions are still specified homogeneously to all countries and regions.

We develop an alternative representation of income distribution data based on principal component analysis (PCA) and test it, along with the lognormal assumption, against a new, consistent dataset we recently constructed (supplementary information S1). We then identify and test a PC-based model and use it, combined with the projected Gini coefficients from Rao *et al* (2019), to produce a new set of global, country-specific income distributions consistent with

the SSPs. We illustrate differences relative to existing lognormal-based projections.

## 2. Data and methods

Development of our projection model required three steps, briefly described here and then each described in more detail below.

1. *Transforming available data to a single dataset.* We developed our own dataset to fit our model since a common issue across all methods is the availability and consistency of data and the metric used to represent income. Some studies represent income using gross domestic product (GDP) per capita (Crespo Cuaresma *et al* 2018, Hughes 2019, Soergel *et al* 2021), others household income (Hallegatte and Rozenberg 2017), and still others use total expenditures (Fujimori *et al* 2020). Some widely used datasets (Povcal, UNU WIDER) provide different metrics for different groups of countries. The metric of income employed will affect the value of the Gini coefficient and must be harmonized with definitions of poverty; for example, the international definition of extreme poverty of \$1.90 per day is based on post-tax household income. In addition, the country and time coverage of different types of data varies.
2. *Carrying out a PCA and evaluating its fit to the data.* We describe the results of our PCA of the transformed data and evaluate its fit. To illustrate the value-added of our model, we compare the fit of our algorithm to that of the lognormal model, which is currently the most prominent model in the literature. We analyze the model fit for the pooled dataset and separately for individual countries.
3. *Developing a PC-based projection model.* We describe our projections model, which uses the PCA results as a basis for generating projections consistent with the inequality storylines of the SSPs.

### 2.1. Data transformations

We developed a more comprehensive and homogeneous set of income distribution data representing net (post-tax) household-level income for 171 countries (SI 1 of the supplementary information describes the input data in more detail). Data on income distributions are available from a number of sources. These data are largely collected from household surveys (WIDER 2008). However, these data represent different measures of income (consumption vs net income) across countries.

We construct a harmonized dataset of income distribution by decile across countries measured consistently as net income. SI 1 of the supplementary information describes the dataset, all input data

and its construction in more detail. This dataset was derived using existing data from the Luxembourg income survey (LIS) (Ravallion 2015), PovCal (Smeeding and Latner 2015) and UNU WIDER (WIDER 2008). Our approach can be summarized in the following steps,

1. We prioritize using net income data from the LIS or PovCal for each country year.
2. Where no data are available from either of these datasets, we use data for net income from PovCal or any other research studies.
3. Where no data on net income is available for a country-year from any source, we impute the net income from consumption using a regression-based approach.
4. We also validate our dataset by ensuring that it contains adequate global coverage (for example, we ensure that the dataset is not made up only of high-income countries).

Our final transformed dataset covers 171 countries over the period 1968–2018.

### 2.2. Fitting the data: principal components vs lognormal

When dealing with data with multiple dimensions or variables, PCA is routinely used to understand broad trends and to reduce dimensions (Bro and Smilde 2014, Karamizadeh *et al* 2020). Given that our income dataset includes ten variables (income deciles) for each country and time period, we carry out a PCA to see whether it can be efficiently represented with fewer dimensions. Note that the PCA combines observations for countries and time periods along one dimension, thus the analysis is conducted on a two-dimensional matrix (country-year  $\times$  deciles). We carried out the analysis on the full data set, but also tested robustness on a subset of the data. Details of the analysis are shown in section S3 of the supplementary information.

The PCA showed that 84% of the variation in the historical data (all observations) on income deciles could be explained by the first component (PC1), indicating that most countries and time periods share a common shape to their income distributions that varies only by a multiplicative constant. The second component (PC2) explained an additional 14% of variation, indicating that when income distributions in specific countries or time periods differ from the general pattern described by PC1, they usually do so in a systematic way. SI table 6 shows the variation and cumulative variation explained by all of the principal components. We did not retain additional components beyond the first two. SI figure 4 shows the shape of the principal components compared to the data and SI table 8 shows the values of the coefficients for the two components. Note that the coefficient values

were consistent when the analysis was run on a subset of the data that retained only half of the variation (SI table 7).

Our representation of the data can therefore be expressed using equation (1):

$$D_{r,t} = a_{r,t}PC1 + b_{r,t}PC2 \quad (1)$$

where  $D$  is a ten-dimensional vector of income shares for all population deciles in region  $r$  at time  $t$ . PC1 and PC2 are the two principal components, also vectors of length 10, and  $a$  and  $b$  are coefficients of the two principal components specific to each region and time.

### 2.3. A PC-based projection model

Using the PC approach as a projection model thus requires us to develop a model for projecting two variables: the two coefficients of the principal components, rather than developing projection models for ten separate income deciles. We first evaluated the distribution of the coefficients derived from the PCA on historical data relative to potential drivers of inequality. We found that the coefficients of the first principal component ( $a$ ) correlate strongly with the Gini co-efficient (SI figure 5;  $R^2 = 0.96$ ). This suggests that this coefficient captures the broad level of inequality represented by the Gini, including its variation across countries and over time. Specifically,  $a$  acts like the Gini coefficient in that an increase in income share in the upper deciles (D10, D9) is associated with decreases in shares along rest of the distribution. We also evaluated the fit of the coefficients of PC1 against multiple independent variables and found that variables such as average educational attainment at different levels (primary, secondary and tertiary) ( $R^2 = 0.43$ ), TFP ( $R^2 = 0.15$ ), and income (GDP per capita) ( $R^2 = 0.23$ ) and combinations of these can be used to reasonably predict the component, although substantially less well than the Gini. SI table 9 shows the  $R$  squared of different combinations of predictors for the coefficients of PC1 and PC2 when models are trained on half the historical data.

Given the strong correlation between the Gini and  $a$ , and the availability of country-specific projections of the Gini coefficient for alternative scenarios (SSPs) over the remainder of the century (Rao *et al* 2019), we introduce a model where the coefficient of PC1 ( $a$ ) is directly projected using the Gini coefficient. The Gini coefficient projections themselves are driven by changes in average educational attainment at the primary, secondary and tertiary level, TFP, trade openness, and policy variables such as public spending on health and education (Rao *et al*) which are therefore implicitly captured in our Gini-driven model.

We estimate the model by regressing historical Gini coefficient data on coefficients of the first principal component determined by the PCA. We do this

by dividing the total dataset into a training dataset, which constitutes half of the total dataset (all observations pre-2004), and a testing dataset. Results from the training dataset are:

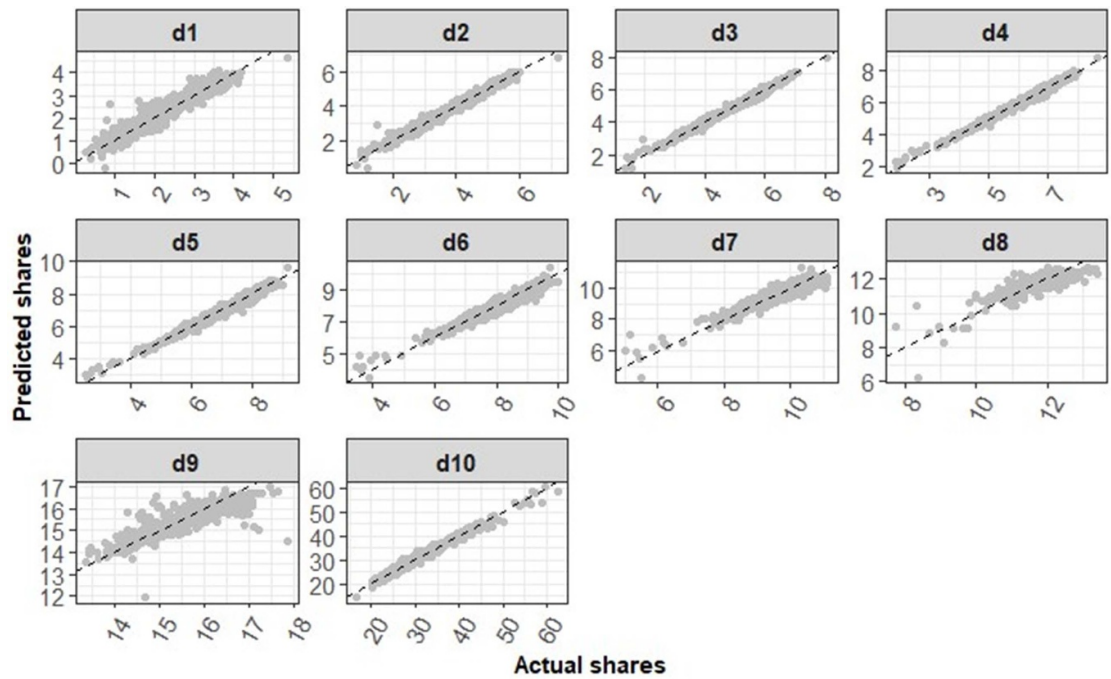
$$a_{r,t} = -11.4815 + 29.71708 \times GINI_{r,t}. \quad (2)$$

Detailed statistics on adjusted  $R^2$  and  $p$ -values are available in SI table 9. The coefficients of the second component ( $b$ ) do not have a significant statistical relationship with any of the variables that share a relationship with  $a$ . This is expected since PC2 by definition is orthogonal to PC1. Examining the loadings matrix for the two components (SI table 7), we find that PC2 acts differently than PC1 by redistributing income shares from deciles at both ends of the distribution (D10, D1) to the higher interior deciles (D5-D9), or vice versa, depending on the sign of its coefficient. We also evaluated the impact of including the second component by examining the change in decile values for different values of the coefficient for PC2 across all countries (SI figure 14). We observe that a high value of the coefficient leads to declines in the shares of d10 and d1 and a redistributive impact on deciles 5–9. For example, a coefficient of 5 leads to a decline in income shares of approximately 5% and 1% for d10 and d1 respectively and increases in shares of income of 2.5% for d9 and d8, and 1% for d7 and d6.

Given our goal of a model for projecting longer-term trends in income distribution rather than temporary fluctuations, we fit a model for the coefficients of PC2 ( $b$ ) using only observations for countries in which PC2 had a consistent impact for at least three consecutive years. We find that the coefficient of the second component has a consistent impact over time in 33 countries (SI table 10) such as the United States while its effect is temporary in other countries such as Malawi (SI figure 6). Since the second component is orthogonal to the first and we are predicting the coefficient of PC1 using the Gini coefficient, we tested independent variables other than the Gini coefficient that might predict the coefficient of the second component. These include lagged measures of the income distribution itself such as the Palma Ratio ( $d10/(d1 + d2 + d3 + d4)$ ) and the share of income in specific deciles, as well as socio-economic drivers such as the labor share of GDP. Note that we use data on the labor share of GDP from the Penn World Tables (version 10) (Feenstra *et al* 2015).

We found that the lagged values of the Palma Ratio and income share in the ninth decile and the current period labor share of GDP were statistically significant in predicting the coefficient of PC2 ( $R^2 = 0.43$ ). Both the labor share of GDP and income share of the ninth decile increase the redistributive impact of PC2. The model, based again on the training dataset, is described in equation (3):





**Figure 1.** Actual income shares (x axis) for the testing dataset vs predicted income shares using the PC projection model estimated on the training dataset (post 2004 values) for each decile (facet). Each facet represents a different income decile. Income shares are expressed as a percent of total income.

$$b_{r,t} = -17.77579 + (0.99944 \times \text{LabShareGDP}_{r,t}) \\ + (112.35374 \times \text{Ninth Decile}_{t-1}) \\ + (-0.34552 \times \text{PalmaRatio}_{r,t-1}). \quad (3)$$

We tested the fit of the projection models for **a** and **b** on a subset of the dataset (figure 1) and find that it yields a generally good fit for all deciles across regions and time periods. Note that the model was fit on a different subset of the dataset. Even though overall fit was found to be reasonable, we find that there is more error for the ninth decile compared to other deciles.

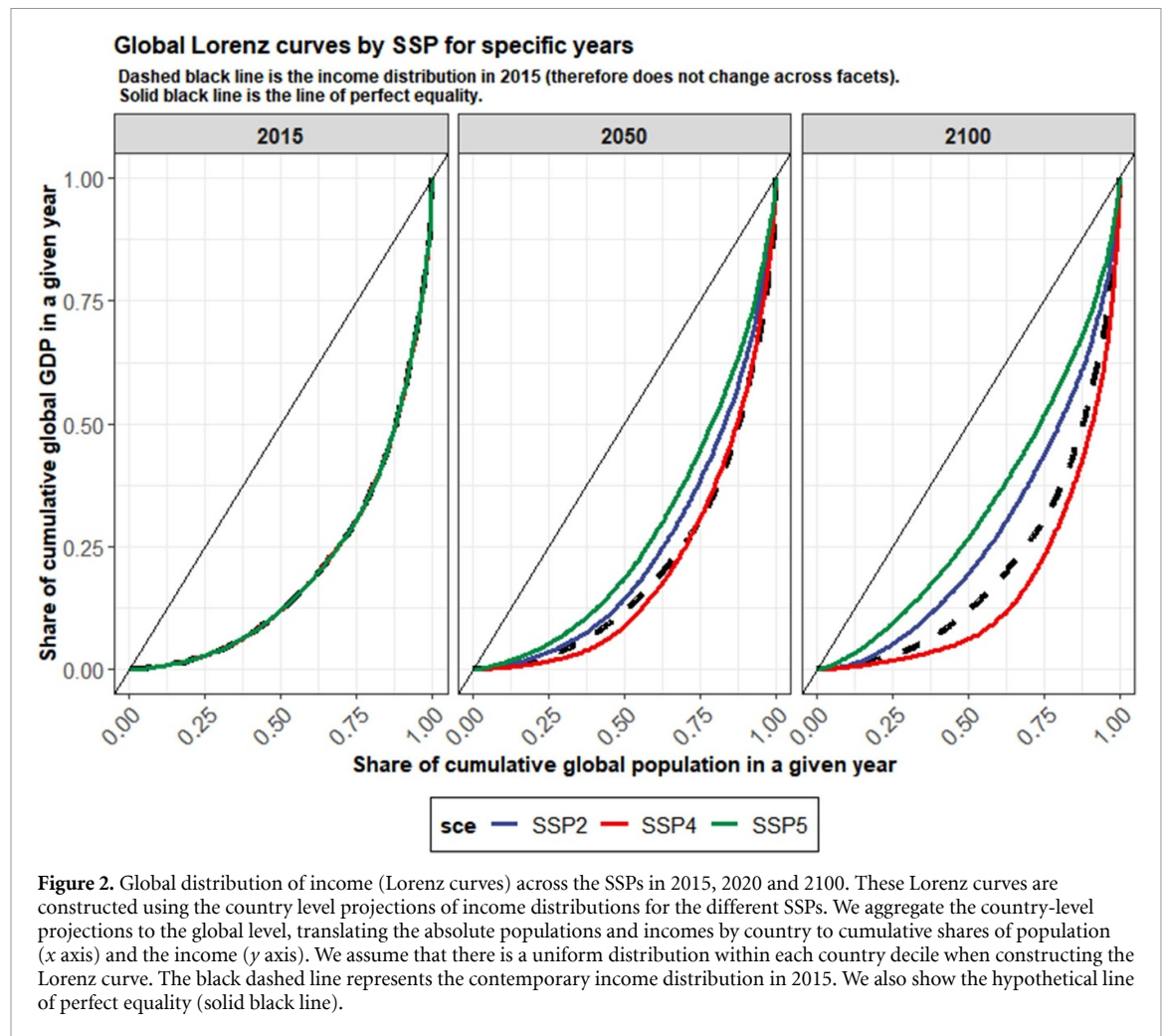
In summary our projection model predicts the coefficient of PC1 (**a**) based on the Gini coefficient and the coefficient of PC2 (**b**) based on a combination of lagged measures of the income distribution and the labor share of GDP. The projected decile values are derived from the equation relating the coefficients and principal components to the income distribution equation (1). To carry out projected future within-country distributions in terms of income shares, we drive the model described above with exogenous Gini coefficients consistent with all five SSPs (Rao *et al* 2019), and hold the labor share of GDP in each country constant, due to a lack of country specific projections of this variable. We compute the lagged value of the Palma ratio and the ninth decile in each time step for use in each time step. When producing results in terms of absolute income and aggregating across countries, we combine projected distributions of income shares with national-level projections of

population (KC and Lutz 2017) and GDP (Dellink *et al* 2017) consistent with the SSPs.

### 3. Results from the projection model

We begin by showing the global distribution of income across selected SSP scenarios as Lorenz curves (figure 2) in different years (2015, 2050, 2100), constructed from the underlying within country distributions. The Lorenz curves compare the cumulative share of the global population with the corresponding cumulative share of income as continuous variables. The Lorenz curve being continuous are constructed assuming a uniform distribution of income within each decile for each country. Consistent with the SSP narratives (O'Neill *et al* 2017), global inequality declines the most in SSP1 and 5 and increases relative to today for SSP 3 and 4, with SSP4 representing the most unequal distribution of income globally. These results depend not only on projected within-country distributions of income shares, but also differences across countries in GDP per capita and population growth.

There is considerable heterogeneity in the country level income distributions within the same SSP scenario in the projections. Consider the income distributions in China and India in 2050 (figure 3). Distributions in China across three SSPs that span the range of inequality do not diverge much by 2050, while in India they diverge substantially. In addition, inequality differs strongly between the two countries



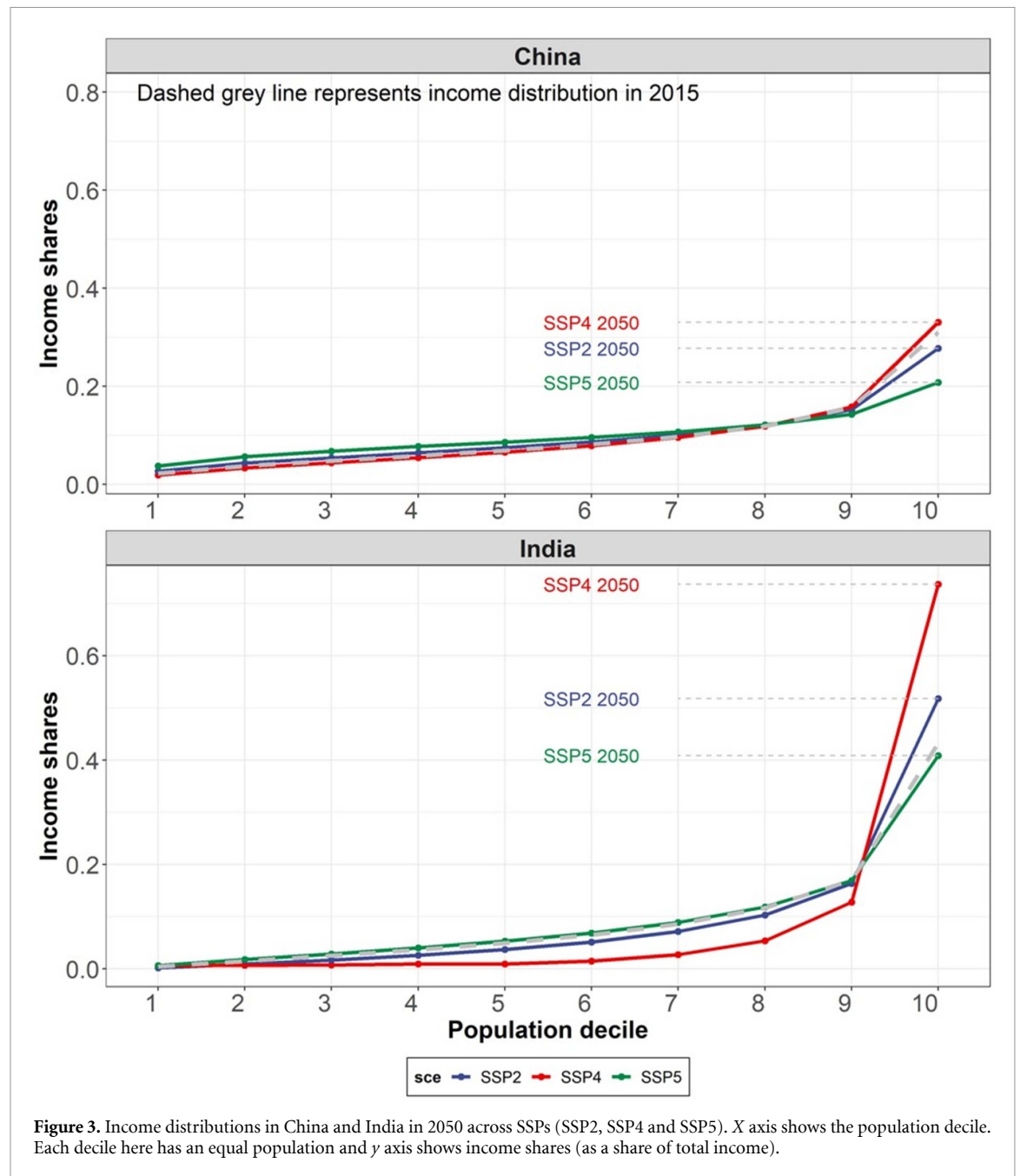
in each SSP, with inequality higher in India. These pathways are driven by the Gini coefficient projections which in turn are driven by different trajectories of TFP and educational attainment consistent with each SSP.

As discussed in the methods section, the PC model generates a better fit to the base data compared to the lognormal approach. We evaluated the fit of the two-component representation to the base data (both in and out of sample). We find that the two-component representation produces a very good fit to the data and is an improvement across all deciles compared to using the Gini and the lognormal functional form (figure 4). In particular, the Gini and lognormal approach produces a nearly consistent underestimate for the tenth decile compared to the PC approach and typically an overestimate for deciles 6–9 and for deciles 4–5 when shares are low. In other deciles, the lognormal errors are not as large, but are still larger than those in the PCA approach.

In addition to the validation for the pooled dataset above, we also conducted country level comparisons (SI figure 12) to confirm that the PC algorithm

provides a better fit to the data at the country level. We also calculated the squared error (data- model)<sup>2</sup> for each country year observation for both the models and compared the distribution of the error across all countries for each decile (SI figure 13). We find that the PC algorithm reduces the mean squared error across all countries by roughly 97% for the 10th decile compared to the lognormal model and similarly reduces the mean squared error by 50% for the 1st decile.

In future years (Our last base year is 2015 and projections range from 2016–2100), this results in higher projected levels of inequality for the same Gini coefficient. We generated projections both for individual countries and 32 aggregated regions (SI figure 3). We find that the PC model generally projects higher levels of income shares in the 10th decile across regions and generally lower values of income shares across other deciles (figure 5, for the example of SSP2 in 2050), although differences are more mixed for the lowest deciles (decile 1–4). The differences in all deciles are also more pronounced in low-income regions which has important implications for demand for



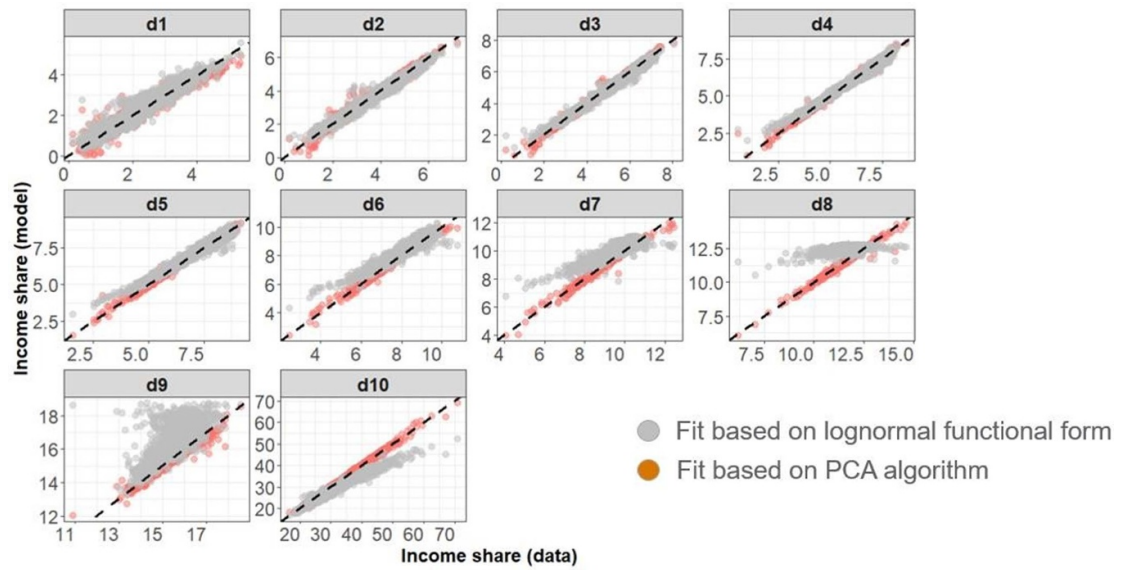
goods and services when using these projections. We find qualitatively similar differences between models across all SSPs (figure 5), with the largest differences from the lognormal in the most unequal SSP (SSP4), with differences as high as 15% for the 10th decile.

While differences in income share projections between models are most prominent for the upper deciles, there are important differences in absolute incomes for the lower deciles (figure 6). We find that the PC model projects changes in income that range from +100% to −50% for the lower deciles, especially for SSPs with high inequality (e.g. SSP4).

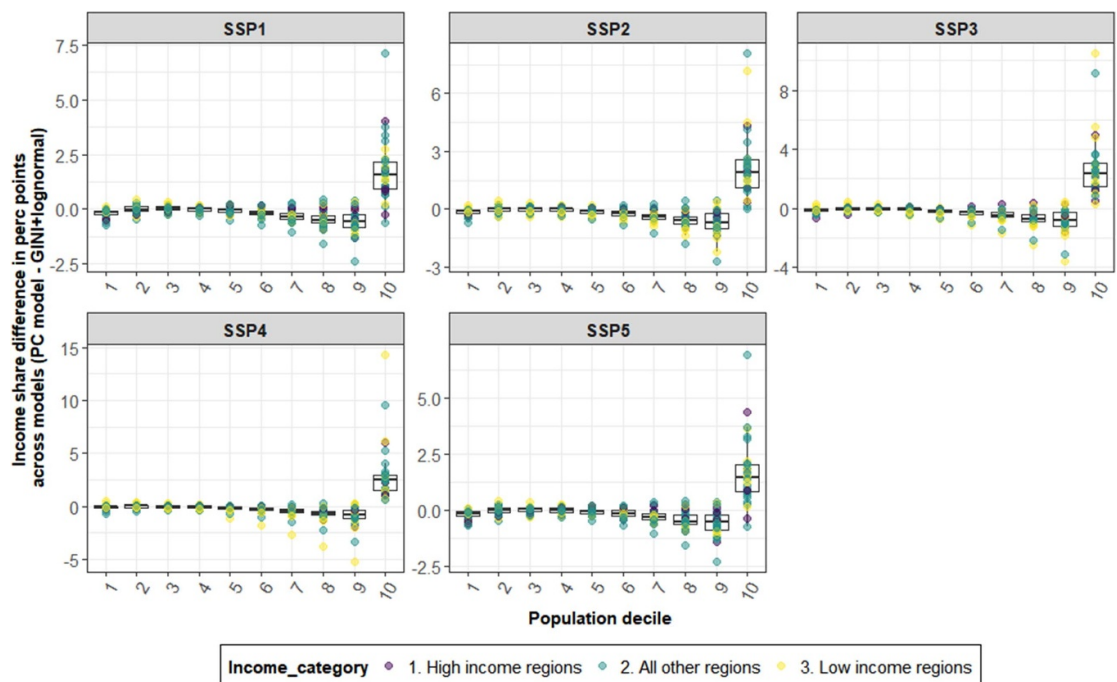
These differences are most prominent in low-income regions which has important implications on consumption of goods and services and may significantly impact poverty estimates.

Since the PC model produces projections of higher inequality, we also compared projections of the Palma ratio which is defined as the ratio of the income of the top decile to the incomes of the bottom four deciles ( $d_{10}/(d_1 + d_2 + d_3 + d_4)$ ). While the Gini coefficient is more representative of the middle of the income distributions, the Palma ratio better reflects the extremes of the distribution. The relationship between the Gini and the Palma ratio is





**Figure 4.** Comparison of fit of lognormal functional form (grey dots) with PCA based fit (orange dots) with data for each decile (facet). Lines represent 1–1 fit between  $x$  and  $y$  axis. Income shares are expressed as a percent of total income.

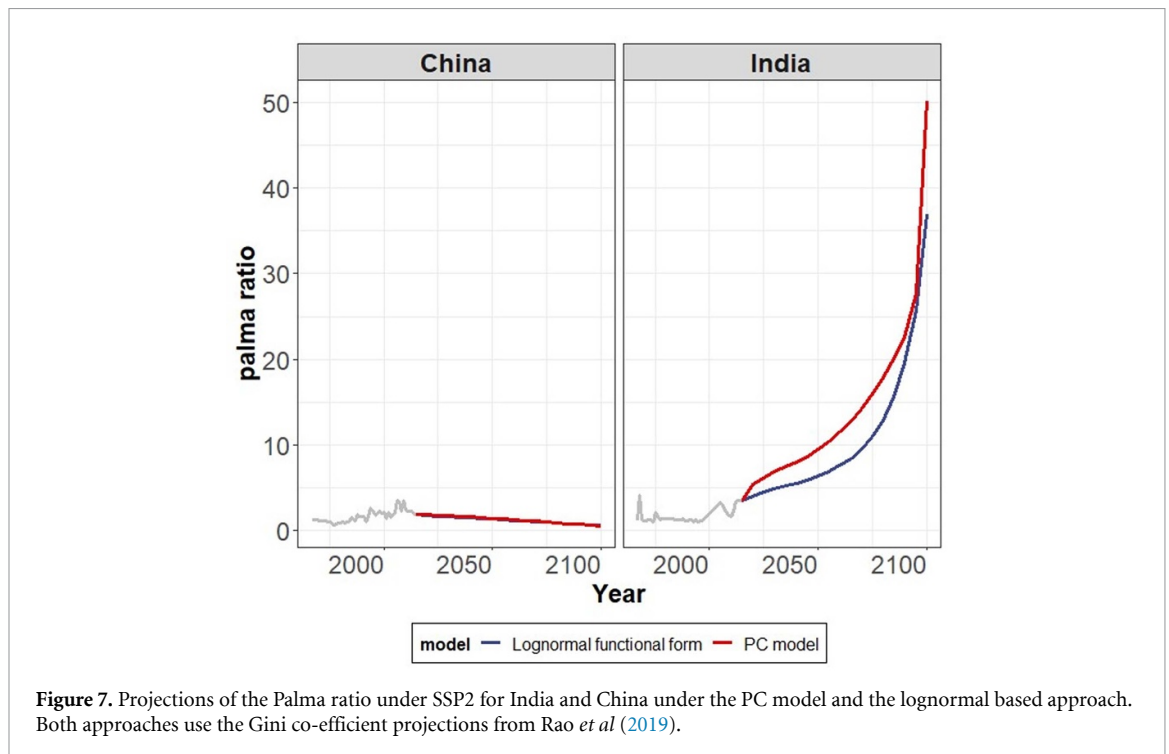
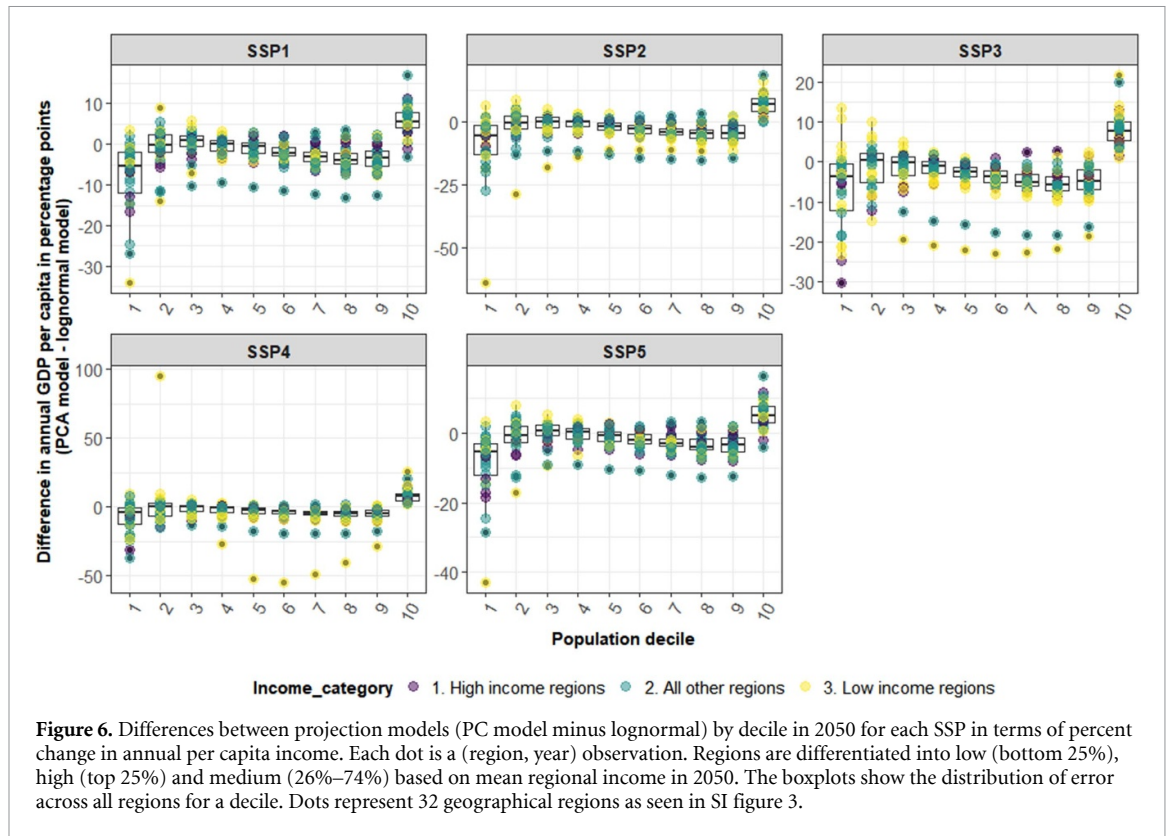


**Figure 5.** Differences in income shares between projection models (PC model minus lognormal) in 2050 for each SSP in terms of share of income. Each dot is a (region, year) observation. Regions are differentiated into low (bottom 25%), high (top 25%) and medium (26%–74%) based on mean regional income in 2050. The boxplots show the distribution of error across all regions for a decile. Dots represent 32 geographical regions as seen in SI figure 3.

approximately exponential (SI figure 7) with small changes in the Gini coefficient corresponding to large changes in the Palma ratio.

We find that the PC model produces projections of higher levels of extreme inequality compared to the lognormal functional form (figure 7 and SI figure 10) across all SSPs, even though both models use

the same Gini coefficient. For example, in SSP2, the Palma ratio in India is projected to rise from a current value of 4 to a value of 50 by 2100 under the PC model compared to 36 using the lognormal approach. In regions where inequality is projected to decline, such as China, the projected Palma ratios are more similar between models.



#### 4. Discussion

In this analysis we created a harmonized dataset of income deciles across countries and used it to develop a model to understand and project within-country income distribution as expressed by deciles. After estimating principal components from

historical data we use existing projections of Gini coefficients to produce future income distributions for countries over the 21st century consistent with the SSPs. These projections can be used to produce different metrics of income distribution and income inequality including global and regional Lorenz curves (which show the cumulative distribution of

income) and the Palma ratio, an indicator of extreme inequality.

Our projections and all input data are available for public use (See link to the Zenodo repository in the data availability statement below).

The PC-based model provides a better fit to data on household income distributions compared to the current predominant method, which combines projected Gini coefficients with the assumption of a lognormal income distribution. The lognormal functional form almost universally underestimates the income shares of the richest populations, and overestimates shares in the poorest decile. This results in the PC model projecting higher levels of inequality in future periods, with more people at the lowest income levels.

These differences are important for several reasons. First, they imply a higher level of vulnerability to climate and environmental impacts than do other income distribution projections, since income and inequality are considered important determinants of that vulnerability. Thus, climate change impacts affecting the poorest populations may be somewhat larger than previously thought. Second, they imply a stronger role for income distribution in driving aggregate consumption; a more unequal distribution of income will have larger effects than a less unequal distribution, assuming demand is non-linear in income. Third, policy responses such as mitigation or adaptation policies that have distributional consequences for the population may have stronger effects than previously estimated, given the potential for higher underlying inequality.

One important point to emphasize is that with the PC based model two countries with similar GINI coefficients can now have different underlying income distributions. This is a dynamic that is missing when simply using the lognormal functional form. It is possible in the PC model because the second component (PC 2) captures dynamics of the income distribution that are not captured by the GINI coefficient. As an example (SI figure 15), the income distributions for Portugal, Italy and Armenia in 2015 (all of which have a GINI of 0.34) are different when using the PC model since the PC2 values for the three countries are different.

There are several possible directions for future work. The PC based model is partially driven by exogenous projections of the Gini coefficient, which are in turn projected using many other variables including income levels (GDP per capita), educational attainment levels (Primary, Secondary and Tertiary), and TFP. The predictors of the PC model could be expanded so that it does not rely on exogenous projections of the Gini coefficient.

In addition, the second principal component within the model is driven in part by labor share

of GDP, which we currently hold constant given the lack of available projections for this variable. We ran various scenarios where we increased the labor share of GDP in order to test the sensitivity of this variable to the current projections (SI figure 11). We found that while the variable is statistically significant, it has small changes on the predicted shares. This aspect of the model can be improved upon in future work.

Similarly, the PC model can be combined with projections of the Gini coefficient other than the one used here to produce projections of income distributions. For example, models that generate their own endogenous projections of Ginis (Hughes 2019) could use the PC model to derive the underlying income distribution.

Finally, the PC model is calibrated to decile level data. If data becomes available on more detailed segments of the income distribution such as the top 1% share or the bottom 1% share, the PC model can be re-calibrated to reproduce these new segments.

## Data availability statement

The main projections described in this paper along with the data used for model development are available here along with code to re-create the main figures- <https://zenodo.org/record/7474549#.ZBsvUcLMJaQ>. Any other data or code can be made available by the authors upon reasonable request.

## Acknowledgments

This research was supported by the U.S. Department of Energy, Office of Science, as part of research in MultiSector Dynamics, Earth and Environmental System Modeling Program. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under Contract DE-AC05-76RL01830. The authors also acknowledge Dr Jon Samperdro, Dr Siwa Msangi at JGCRI, PNNL and Dr Barry Hughes, Dr Mohammad T Irrfan at the University of Denver for their input and feedback on this work.

## ORCID iDs

Kanishka B Narayan  <https://orcid.org/0000-0001-8483-6216>

Brian C O'Neill  <https://orcid.org/0000-0001-7505-8897>

Stephanie T Waldhoff  <https://orcid.org/0000-0002-8073-0868>

Claudia Tebaldi  <https://orcid.org/0000-0001-9233-8903>

## References

- Arnold B C 2014 Pareto distribution *Wiley StatsRef: Statistics Reference Online* pp 1–10
- Badel A, Huggett M and Luo W 2020 Taxing top earners: a human capital perspective *Econ. J.* **130** 1200–25
- Bourguignon F 2015 The globalization of inequality *The Globalization of Inequality* (Princeton, NJ: Princeton University Press)
- Bro R and Smilde A K 2014 Principal component analysis *Anal. Methods* **6** 28122831
- Campagnolo L and Davide M 2019 Can the Paris deal boost SDGs achievement? An assessment of climate mitigation co-benefits or side-effects on poverty and inequality *World Dev.* **122** 96–109
- Chen J, Xian Q, Zhou J and Li D 2020 Impact of income inequality on CO<sub>2</sub> emissions in G20 countries *J. Environ. Manage.* **271** 110987
- Chitiga M, Sekyere E and Tsoanamsie N 2015 Income inequality and limitations of the Gini index: the case of South Africa (Human Sciences Research Council (HSRC)) (available at: [www.hsrc.ac.za/en/review/hsrc-review-november-2014/limitations-of-gini-index](http://www.hsrc.ac.za/en/review/hsrc-review-november-2014/limitations-of-gini-index)), site accessed, 2
- Chotikapanich D 2008 *Modeling Income Distributions and Lorenz Curves* vol 5 (Springer Science & Business Media)
- Crespo Cuaresma J, Fengler W, Kharas H, Bekhtiar K, Brottrager M and Hofer M 2018 Will the sustainable development goals be fulfilled? Assessing present and future global poverty *Palgrave Commun.* **4** 29
- Dellink R, Chateau J, Lanzi E and Magné B 2017 Long-term economic growth projections in the Shared Socioeconomic Pathways *Glob. Environ. Change* **42** 200–14
- Feenstra R C, Inklaar R and Timmer M P 2015 The next generation of the Penn world table *Am. Econ. Rev.* **105** 3150–82
- Figueira F C, Moura Jr N J and Ribeiro M B 2011 The Gompertz–Pareto income distribution *Physica A* **390** 689–98
- Fosu A K 2017 Growth, inequality, and poverty reduction in developing countries: recent global evidence *Res. Econ.* **71** 306–36
- Fujimori S, Hasegawa T and Oshiro K 2020 An assessment of the potential of using carbon tax revenue to tackle poverty *Environ. Res. Lett.* **15** 114063
- Hallegatte S and Rozenberg J 2017 Climate change through a poverty lens *Nat. Clim. Change* **7** 250–6
- Hallinan Jr A J 1993 A review of the Weibull distribution *J. Qual. Technol.* **25** 85–93
- Hertel T W, Burke M B and Lobell D B 2010 The poverty implications of climate induced crop yield changes by 2030 *Glob. Environ. Change* **20** 577–85
- Hughes B B 2019 *International Futures: Building and Using Global Models* (New York: Academic)
- Hussein Z, Hertel T and Golub A 2013 Climate change mitigation policies and poverty in developing countries *Environ. Res. Lett.* **8** 035009
- Jafino B A, Walsh B, Rozenberg J and Hallegatte S 2020 Revised estimates of the impact of climate change on extreme poverty by 2030
- Karamizadeh S, Abdullah S M, Manaf A A, Zamani M and Hooman A 2020 An overview of principal component analysis *J. Signal Process. Syst.* **4** 173
- Liu C, Jiang Y and Xie R 2019 Does income inequality facilitate carbon emission reduction in the US? *J. Clean. Prod.* **217** 380–7
- Mandelbrot B 1960 The Pareto–Levy law and the distribution of income *Int. Econ. Rev.* **1** 79–106
- O'Neill B C, Kriegler E, Ebi K L, Kemp–Benedict E, Riahi K, Rothman D S, van Ruijven B J, van Vuuren D P, Birkmann J and Kok K 2017 The roads ahead: narratives for shared socioeconomic pathways describing world futures in the 21st century *Glob. Environ. Change* **42** 169–80
- O'Neill B C, Kriegler E, Riahi K, Ebi K L, Hallegatte S, Carter T R and Van Vuuren D P 2014 A new scenario framework for climate change research: the concept of shared socioeconomic pathways *Clim. Change* **122** 387–400
- Osberg L 2017 On the limitations of some current usages of the Gini index *Rev. Income Wealth* **63** 574–84
- Rao N D and Min J 2018 Less global inequality can improve climate outcomes *Wiley Interdiscip. Rev.: Clim. Change* **9** e513
- Rao N D, Sauer P, Gidden M and Riahi K 2019 Income inequality projections for the shared socioeconomic pathways (SSPs) *Futures* **105** 27–39
- Ravallion M 2001 Growth, inequality and poverty: looking beyond averages *World Dev.* **29** 1803–15
- Ravallion M 2015 The Luxembourg income study *J. Econ. Inequal.* **13** 527–47
- Riahi K, Van Vuuren D P, Kriegler E, Edmonds J, O'Neill B C, Fujimori S and Tavoni M 2017 The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: an overview *Glob. Environ. Change* **42** 153–68
- Rinne H 2008 *The Weibull Distribution: A Handbook* (Chapman and Hall/CRC)
- Samir K C and Lutz W 2017 The human core of the shared socioeconomic pathways: population scenarios by age, sex and level of education for all countries to 2100 *Glob. Environ. Change* **42** 181–92
- Sampedro J, Iyer G, Msangi S, Waldhoff S, Hejazi M and Edmonds J A 2022 Implications of different income distributions for future residential energy demand in the US *Environ. Res. Lett.* **17** 014031
- Smeeding T and Latner J P 2015 PovcalNet, WDI and 'All the Ginis': a critical review *J. Econ. Inequal.* **13** 603–28
- Soergel B, Kriegler E, Bodirsky B L, Bauer N, Leimbach M and Popp A 2021 Combining ambitious climate policies with efforts to eradicate poverty *Nat. Commun.* **12** 2342
- Van der Mensbrugghe D 2015 Shared socio-economic pathways and global income distribution *GTAP paper*
- Wider UNU 2008 *World Income Inequality Database* (Helsinki: WIDER)