



OPEN

DATA DESCRIPTOR

A dataset of income distribution on provincial, urban, and rural levels for China from 2020 to 2100

Mingyu Lei^{1,2}, Setu Pelz³, Shonali Pachauri³ & Wenjia Cai^{1,2}✉

Projections of future income distributions at subnational levels are becoming increasingly important for a variety of analyses and evaluations. However, relevant datasets are currently limited. This study presents a methodological framework that introduces machine learning algorithms to a top-down approach used for generating income distribution datasets. We project per capita disposable income and income inequality for 31 Chinese provinces from 2020 to 2100, considering different scenarios based on China's local circumstances, and then estimate income distributions based on these. After accounting for necessary consistency between provincial, urban, and rural income datasets, we further generate the same data products at the urban and rural level for each province. We validate our projection results drawing on data from 2007–2023 for China's disposable income, data from 2007 to 2019 for provincial income inequality in China, as well as national income inequality data for the past 20 to 60 years from select developed countries. The proposed methodology provides flexibility to generate similar data products according to a user's specific needs. Our resulting datasets have several potential applications and can serve as inputs for research on drivers and impacts across social, economic, and environmental domains.

Background & Summary

Projections of income distribution are becoming increasingly important for various research purposes. Income distribution is a significant factor in determining consumption and social wellbeing, as well as their uneven distribution among populations. It also closely relates to the ability of diverse populations to cope and adapt to anticipated or unexpected stressors. Scientific projections of income distribution are firstly, essential for conducting scenario analyses relevant to many important societal, economic, and environmental issues, including, but not limited to, demand assessments for a variety of commodities such as energy, water, food, and land use^{1,2}, estimating environmental footprints^{3,4}, cost-benefit evaluation of policies^{5,6}, and impacts, adaptation, and vulnerability (IAV) related to climate change and other disasters^{7,8}. In addition, income distribution projections provide an opportunity to reveal the considerable differences among populations hidden in the current aggregated national results. The use of such projections is gaining importance in multi-objectives scenario studies^{9,10}, particularly in aligning across multiple Sustainable Development Goals (SDGs) such as poverty eradication and climate mitigation^{11,12}.

The need for income projections is becoming more prominent, but current research and methodologies to support this are limited. Previous literature includes attempts to project future income distribution considering specific metrics such as GDP¹³, income inequality¹⁴, poverty rate¹⁵, or income level by deciles¹⁶. Our interest, however, is to project full income distributions. To this end, two broad approaches might be feasible, as outlined in Table 1. The top-down approach is the most commonly used method, which relies on existing projections of per capita disposable income, income inequality (measured by Gini coefficients)¹⁴ and a specific assumed form of income distribution, such as the log-normal distribution^{17,18}, Weibull distribution, or an emerging non-parametric distribution^{16,19}. An alternative approach is microsimulation, a bottom-up method that uses a large amount of individual/household survey data and assumptions about the dynamics of socio-demographic

¹Department of Earth System Science, Institute for Global Change Studies, Ministry of Education Key Laboratory for Earth System Modeling, Tsinghua University, Beijing, 100084, China. ²Tsinghua-Rio Tinto Joint Research Centre for Resources, Energy and Sustainable Development, International Joint Laboratory on Low Carbon Clean Energy Innovation, Laboratory for Low Carbon Energy, Tsinghua University, Beijing, 100084, China. ³International Institute for Applied Systems Analysis, Laxenburg, Austria. ✉e-mail: wcai@tsinghua.edu.cn

Methodology	Literature	Spatial level	Future inequality	Functional form
Top-down: Income inequality + particular functional form	Calzadilla ⁵¹	Global	Constant	Lognormal
	Rao <i>et al.</i> ¹⁴	Country	Regression-based	
	Fujimori <i>et al.</i> ¹⁷	Country	Rao <i>et al.</i> ¹⁴	Lognormal
	Soergel <i>et al.</i> ¹⁸	Country	Rao <i>et al.</i> ¹⁴	Lognormal
	Narayan <i>et al.</i> ¹⁶	Country	Rao <i>et al.</i> ¹⁴	Non-parametric
	Casper <i>et al.</i> ¹⁹	U.S. State	Rao <i>et al.</i> ¹⁴	Non-parametric
Bottom-up: Microsimulation	Hallegatte and Mauro ²⁰	Global		
	Hallegatte and Rozenberg ⁷	Country		

Table 1. Overview of key previous studies on income distribution projections.

characteristics for a set of representative households^{7,20}. The bottom-up approach has limitations in its application to nations and regions where access to the required survey data is less possible. In contrast, the top-down approach's ability to generalize makes it easier to adopt at different spatial levels, as demonstrated by previous literature referenced in Table 1.

Despite the viability of the methodology for performing subnational projections, previous work on projecting income distributions is still heavily limited to the national-level, which does not fully support detailed sub-national analyses²¹. The foremost reason is the absence of income inequality datasets at sub-national levels, which has resulted in reliance on a published national-level dataset of Gini coefficient projections¹⁴. A recent study attempted to generate income distribution projections at the U.S. state-level based on this dataset¹⁹, but it assumed that the state-level Gini coefficients would follow the same growth rate as at the national level. This assumption undermines the heterogeneity in income distributions across states, even after accounting for the varying base year Gini coefficients of states. Additionally, previous studies usually use projections of GDP per capita as a proxy for future disposable income^{16,19}, which inevitably leads to an overestimation of income, as GDP per capita is typically higher than household disposable income. Robust projections of disposable income and income inequality are indispensable to forecast income distribution at sub-national levels. Traditional econometric methods often used in previous studies, however, are not always reliable for making such long-term projections¹⁴. Machine learning (ML) algorithms offer an alternative to traditional econometric methods²², and have been applied to predict future socioeconomic conditions using indicators such as population²³, energy demand^{24,25}, price indices²⁶, and consumption behaviours²⁷.

Previous research has shown that the lack of subnational projections on income distributions is mainly due to the absence of scientific long-term projections for disposable income and income inequality, as well as a systematic framework for projecting them. Therefore, this study aims to address two sub-tasks. First, following the top-down approach, we develop a methodological framework using ML algorithms to generate income datasets of provinces based on their diverse characteristics. Then, using this approach, we project per capita disposable income, income inequality (measured by Gini coefficients), and income distributions for 31 Chinese provinces from 2020 to 2100, considering different scenarios based on China's local circumstances. The primary data product we generate is provincial projections. Additionally, considering necessary consistency constraints between provincial, urban, and rural income datasets, for each province, we also provide results at urban and rural level as a subsidiary dataset. The focus is on China due to its growing global significance and the huge diversity among Chinese provinces, that allows for assessing the methodology's effectiveness in capturing heterogeneities across provinces.

Methods

Model design under consistency constraints. Our methodological framework mainly consists of a provincial model and urban (rural) model, as shown in Fig. 1. Each model is composed of a training and simulating module and is expected to deliver three datasets at corresponding spatial level, including per capita disposable income ($PD_1, SD_{1-1,1-2}$), income inequality measured by Gini coefficient ($PD_2, SD_{2-1,2-2}$), and income distribution ($PD_3, SD_{3-1,3-2}$). These datasets cannot be generated separately because they are not independent of each other but subject to a number of qualitative or quantitative consistency constraints.

For disposable income, projections of provincial income are expected to keep consistent with future economic development (see e.g. this published GDP dataset¹³). Meanwhile, provincial, urban, and rural income need to be consistent (Eq. 1), such that the projected provincial income should equal population-weighted averages of urban and rural income. In terms of Gini coefficients, a proxy of income inequality, the consistency constraint for projections of provincial, urban, and rural Gini coefficients is described as Eq. 2²⁸. The income distributions at provincial, urban, and rural level are then generated based on the predicted per capita disposable income and Gini coefficients. The relationships between disposable income, Gini coefficients, and income distributions suggest that the factors considered for training and simulating should be derived keeping the relationships in mind and can bridge the three outcomes we require, so that we can solve for outcomes by combining the constraints and the response factors, rather than predicting them separately.

$$I = PS_{ur} \times I_{ur} + PS_{ru} \times I_{ru} \quad (1)$$

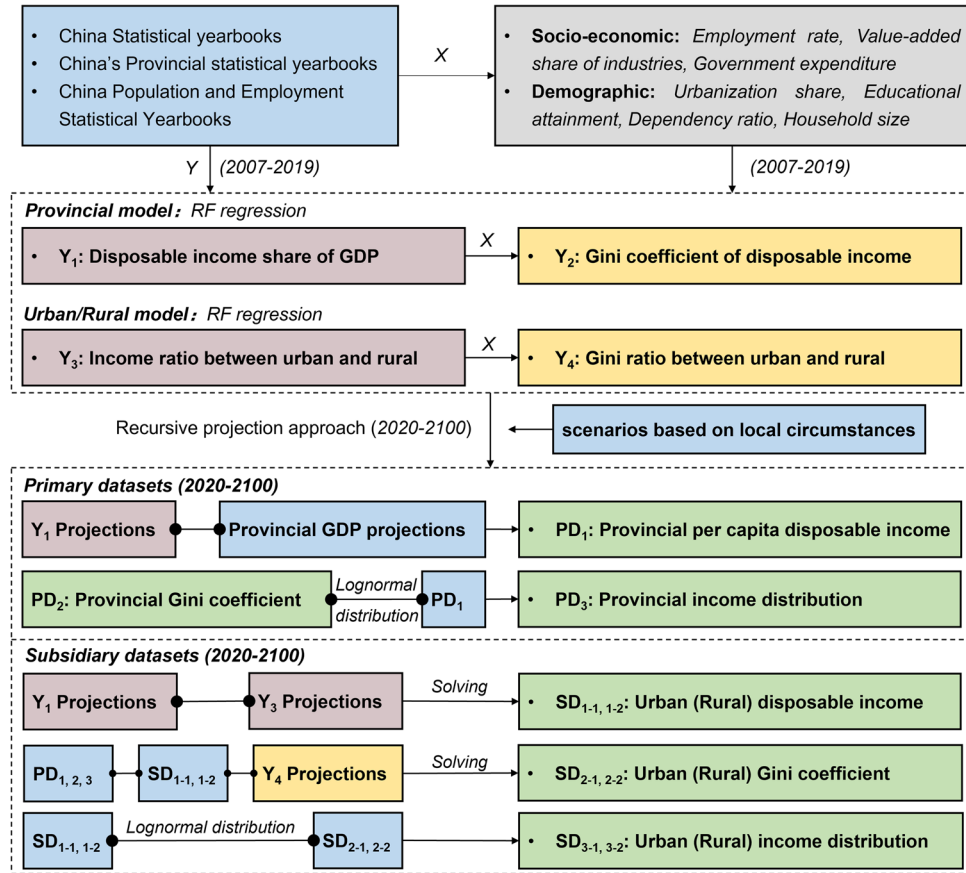


Fig. 1 Methodology of projecting disposable income, income inequality and income distribution.

$$ProGini = PS_{ur}^2 \times \frac{I_{ur}}{I} \times UrGini + PS_{ru}^2 \times \frac{I_{ru}}{I} \times RuGini + PS_{ur} \times PS_{ru} \times \frac{I_{ur} - I_{ru}}{I} \quad (2)$$

Where PS_{ur} and PS_{ru} represent the urban and rural population share, respectively, while I_{ur} , I_{ru} , and I are the per capita disposable income of urban, rural, and the whole province.

For the provincial model, the share of disposable income in GDP (Y_1) and provincial Gini coefficients (Y_2) are selected as factors, and the ratio between urban and rural income (Y_3) and between urban and rural Gini coefficients (Y_4) are chosen for the urban (rural) model. The consistency between projected provincial income and GDP is ensured by combining the factors of the provincial model with the published GDP dataset¹³, while the consistency between provincial, urban, and rural results is guaranteed by solving the corresponding consistency constraint for predicted urban to rural income ratio or Gini ratio.

Data acquisition and processing. Constructing and predicting the selected response factors for provincial and urban (rural) model of 31 Chinese provinces requires a range of datasets at different spatial scales (details can be found in Table 2).

Household disposable income and income inequality. For the period 2007–2019, we first collect the provincial, urban, and rural per capita disposable income and GDP of 31 Chinese provinces from China's Provincial statistical yearbooks. Then, we estimate income Gini coefficients²⁹ of 31 Chinese provinces at provincial, urban, and rural level.

To this end, we first collected grouped household-survey data at urban and rural level from China's Provincial statistical yearbooks. For each urban and rural income group, we used the following indicators to compute Gini coefficients - households surveyed (HN), average household size (HS), average annual per capita disposable income ($PCDI$). For province i at year t , the income Gini coefficients of urban ($UrGini$) and rural ($RuGini$) populations are calculated using Eq. 3. Based on $UrGini$ and $RuGini$ of each province, provincial income Gini coefficients ($ProGini$) for province i at year t are calculated using Eq. 2.

$$Gini = 1 - 1/(P \times W) \sum_{j=1}^n [(W_{j-1} + W_j) \times P_j] \quad (3)$$

Usage	Data	Source	Spatial level			
Gini coefficients	Households surveyed	China's Provincial statistical yearbooks China Population and Employment Statistical Yearbooks	Urban/Rural			
	Household size					
	Per capita disposable income					
Predictive variables	Value-added of industries	Secondary	China Statistical yearbooks	Provincial		
		Tertiary				
	Government spending	Technology				
		Social protection				
		Health				
		Education				
	Employment rate	China's Provincial statistical yearbooks China Population and Employment Statistical Yearbooks			Provincial/Urban/Rural/ Rural	
	Educational attainment					Primary (aged 6–12)
						Secondary (aged 13–18)
						High (aged 19+)
	Dependency structure					Juvenile and children (aged 0–14)
Aged (aged 65+)						
Household size						
Urbanization						

Table 2. Dataset and variables used for establishing provincial and urban (rural) model.

P_j represents the population of urban or rural income-group j , which was obtained by multiplying HN and HS of income-group j , and P is the sum of P_j . W is the cumulative income of P , as measured by the sum of the products of P_j and $PCDI$ of all income-groups, while W_j is the total income accumulated to income-group j .

Due to incomplete or missing data for some provinces for certain years, we also performed a series of data cleaning processes, as detailed in Tables S1–2. For example, for a few provinces, such as Guangxi between 2014 and 2019 and Chongqing between 2013 and 2015, HS data for each income group was missing, so we assumed the same HS across all income groups. Some provinces reported neither HN data nor the criteria used for dividing income groups. In this case, we set the HN of these provinces based on data for years with complete data records.

Socioeconomic and demographic variable selection. Changes in socioeconomic and demographic characteristics are understood to be related to changes in income distributions. Regarding socioeconomic features, several studies indicate that industrial structure^{30,31}, technological progress³², employment rate^{33,34}, and government expenditure^{35,36} are related to household income. For demographic features, urbanization rate, education attainment, household size, and dependency rates are shown to be related to household income^{37,38}.

To capture changes in historical response factors, we selected a wide range of predictive variables (details in Table 2). Specifically, to reflect socioeconomic status of 31 Chinese provinces, we selected the share of value-added of industries in GDP, employment rate, and government spending on various items (including health, education, social protection, and technology), which were collected from China Statistical yearbooks. For demographic factors, we selected educational attainment (four categories: illiterate, primary, secondary, and high level), juvenile and child (J&C) dependency ratio, aged dependency ratio, average household size, and urbanization, which we retrieved from China's Provincial statistical yearbooks and China Population and Employment Statistical Yearbooks. Notably, employment rate, household size, educational attainment, and dependency structure were collected at both provincial, urban, and rural level, while the data on other variables, was only available at the provincial level.

Modelling framework for disposable income and income inequality. This module attempted to build a general modelling framework suitable for both balanced and unbalanced panel data. Using a machine learning framework, we utilized the random forest (RF) regression algorithm to create a data-driven workflow, as shown in Fig. 2. The workflow comprised five steps, including data splitting, key feature selection, hyperparameters optimization, model comparison and baseline validation, and an additional robust validation for unbalanced data.

Dataset construction and splitting. Changes in Gini coefficients might be captured by socioeconomic and demographic variables relating to both current and past years¹⁴. Therefore, for both the provincial, and the urban, and rural level, three datasets were constructed, namely No lags (NL), First-order lag (FL), and First-order lag only (FLO) that contained information on variables considering different time periods.

To apply the RF algorithm, we need to split the dataset into a training and test set. This helps to avoid overfitting and allows us to test the predictive capability of the model. The traditional method of data splitting is sufficiently well suited for balanced panel data such as the dataset of per capital disposable income at both provincial, urban, and rural level. However, the Gini coefficient datasets for provinces, urban and rural areas were

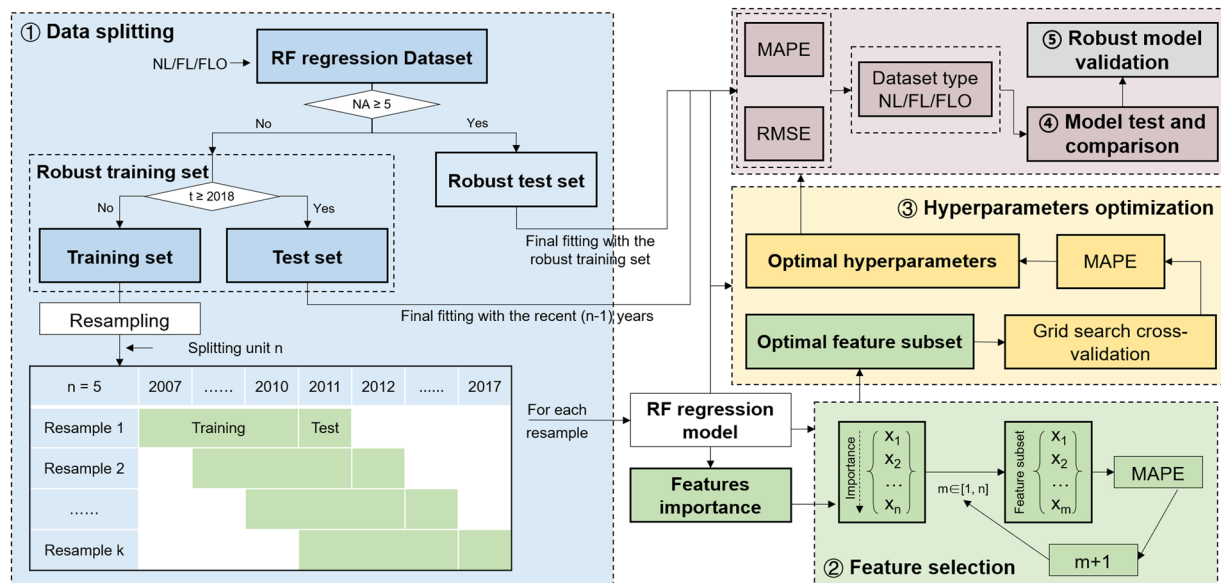


Fig. 2 Workflow of modelling disposable income and income inequality.

unbalanced panel data due to some provinces having incomplete or undisclosed records in certain years. This meant we had random missing datapoints, and as a result, we had to predict the Gini coefficient of province i at year t based on data from other provinces. This required the model to generalize well. So, the RF model needed to perform well both over time and across different locations. For this purpose, we split the Gini coefficients dataset into two separate sections, including a baseline set and a robust test set, as shown in Fig. 2. To build the robust set, we selected a small number of provinces, which were not used to train and test the model, while the remaining provinces were assigned to the baseline set.

For both income dataset and Gini coefficient dataset, we used the data from 2018 and 2019 for the test dataset and data from 2007 to 2017 as the training dataset. The model was first trained on the training set and tested on the test set, to ensure satisfactory performance on the time dimension. Subsequently, for the RF model of Gini coefficients, a more rigorous validation based on the spatial dimension was conducted on the robust set to evaluate the predictive capability of the trained model in predicting the Gini coefficients of provinces that it had not been trained on previously.

During the training process, a time series resampling method was used on the training dataset to create multiple resamples. Each sample was generated by splitting the data into 5-year intervals and then moving forward in 1-year steps, as shown in Fig. 2. In each resample, the first four years of data were used for training the model, and the model was then evaluated using the data from the last year. This approach assured that the model was not trained on later data and then used for predicting earlier data, and it also enhanced the model's ability to generalize across the temporal dimension.

Key features selection. This study used the literature review to inform the selection of several socioeconomic and demographic predictors that are considered closely related to income metrics. However, it is important to empirically determine the optimal subset of predictive factors. Specifically, excluding irrelevant or redundant predictive factors through key features selection is useful not only for preventing overfitting but also for improving the generalizability of the model. This can help in achieving a better prediction performance, as detailed in Fig. 2.

Under the RF framework, we first calculated the importance of each variable, measured by the percentage increase in mean square error (%MSE), on every resample. This helped us to test the capacity of each feature in predicting response factors across multiple time windows. Then, we calculated the average %MSE of each feature across all resamples. We used a forward search approach to explore all feature combinations from the most to least important. For each combination, we fitted the RF model on every resample and calculated the resample-average of the mean absolute percentage error (MAPE) to evaluate the model's performance. This process helped us to identify the optimal feature subset.

Hyperparameters optimization. We used two hyperparameters of the RF model, i.e., n_{tree} and m_{try} , for the model training process. We performed a grid search cross-validation. Specifically, we built a hyperparameter basket and applied it to each resample. We then ran the RF model iteratively on every resample using each parameter combination in the basket. Similar to the method applied for the features selection, we chose MAPE as the performance index and calculated an average of it across all resamples to evaluate the parameter combination and the optimal parameters.

Features	Future development scenarios				Quantitative method/Data source	Spatial level		
	HSD	HQD	BAU	LSD				
Secondary industries	High	Middle high	Medium	Low	Jing, <i>et al.</i> ⁴²	Provincial		
Tertiary industries	Middle high	High	Medium	Low				
GDP	High	Middle high	Medium	Low				
Employment rate	High	High	Medium	Low	Growth rate	Provincial/Urban/Rural		
Household size	Small	Small	Medium	Large	Headship rate			
Primary education	Low	Low	Medium	High	Chen, <i>et al.</i> ⁴³			
Secondary education	High	High	Medium	Low				
High education	High	High	Medium	Low				
J&C dependency ratio	Low	Low	Medium	High				
Aged dependency ratio	High	High	Medium	Low				
Urbanization rate	High	High	Medium	Low				
Technology spending	Regression-based simulation						A regression model ¹⁴	Provincial
Social protection spending								
Health spending								
Education spending								

Table 3. Future assumptions regarding quantitative variables under different scenarios.

Model comparison and baseline validation. After training the model using the optimal feature subset and parameters, we used the test dataset to validate the model's predictive capacity. In addition to the MAPE, we also calculated the root mean square error (RMSE) to assess the model's performance in predicting future Gini coefficients. The MAPE and RMSE were estimated using Eq. 4a,b.

$$MAPE = \frac{100\%}{n} \times \sum_{k=1}^n \left| \frac{Pred_k - Real_k}{Real_k} \right| \quad (4a)$$

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{k=1}^n (Pred_k - Real_k)^2} \quad (4b)$$

Where k represents the number of datapoints included in the test dataset, and $Pred_k$ and $Real_k$ are prediction and real value of response factors for datapoint k (province i , year t) respectively, while $Real$ denotes the average real value of response factors of all datapoints.

For the provincial, urban, and rural level, we trained three RF models on the three datasets (i.e., NL, FL, FLO) and these were evaluated and compared based on MAPE and RMSE to select the model with the best predictive capacity.

Robust validation. For the optimal model of Gini coefficients selected at provincial, urban, and rural level, we then carried out a robust test to assess the generalizability of the model on the spatial dimension. Specifically, we fitted a RF model on the baseline set using the optimal feature subset and parameters trained before, and then tested this on the robust set. The robust test guaranteed the RF model with satisfactory performance in the temporal dimension (predicting a province's future via its historical data) can also perform well on the spatial dimension (predict a province's future via other provinces' historical data).

Future assumptions under different scenarios. *Description of different development scenarios.* We developed four scenarios to describe future development of the 31 Chinese provinces with consideration of their local context, namely the high-speed development (HSD) pathway, high-quality development (HQD) pathway, business-as-usual (BAU) pathway, and the low-speed development (LSD) pathway.

We define HSD to represent an industrialized development pathway with the fastest assumed economic growth rate and characterised by a demographic future of high educational attainment and aging. We describe HQD as a high-quality economic development future. High-quality development represents a pathway that China plans to achieve, and it means shifting the growth model from crude to intensive, with a focus on innovation. In this case, the tertiary industries will play a more important role in the national economy than the secondary industries, while inevitably, some economic growth may be sacrificed. Hence, compared to HSD, we assume a slightly lower economic growth rate in HQD but similar demographic assumptions. We assume the BAU pathway follows historical development trends with moderate changes in socioeconomic and demographic characteristics. Finally, for LSD, we assume a future that is the exact opposite of HSD. The detailed assumptions for each variable in the four scenarios are shown in Table 3.

Quantifying assumptions of predictors under different scenarios. We applied various quantitative methods to define the future values of key variables, as shown in Table 3. We first quantified the variables at the

provincial level. We used predictions from two available datasets. We sourced projections of GDP and the share of value-added of industries from Jing, *et al.*¹³, and of educational attainment, urbanization, and household dependency from Chen, *et al.*³⁹ Those two studies developed localized SSP storylines for China, which allow for consistent assumptions across the two datasets. We then mapped the localized SSP narratives from these two studies to our four scenarios, assuming similar demographic and economic developments as under SSP5, SSP1, SSP2, SSP3 in the HSD, HQD, BAU, and LSD pathways, respectively.

We used past growth rates to generate future employment rate trends. Under HSD and HQD, we assumed an increase in employment at the average rate of increase in employment in G7 countries over the last twenty years, which is about 0.1 percentage per year. Under BAU, we assumed the employment rate increases at the rate of 0.05 percentage per year, while under LSD we assumed the employment rate to stay at the level it was in 2019. We adopted the headship rate method^{40,41} to produce household size projections, based on data from the Chinese Census 2000 and 2010, and the provincial projections of population and urbanization rate³⁹.

We did not have access to projections or commonly used quantitative methods for predicting government spending. We therefore developed and applied a regression model to create a regression-based simulation for future government spending on four specific items¹⁴. This model estimated the spending of each item using a combination of socioeconomic and demographic variables (in a first-order lag form), along with available future projections. We based our model on provincial panel data from 2007 to 2019, and included province fixed effects and a time variable (Year). The performance of the regression model can be seen in Figs. S1–4.

It is important to note that the projections of variables were done at the same spatial level as the available historical data. As a result, we projected value-added of industries, urbanization, and government spending at the provincial level. Household size and employment rate were calculated based on the respective historical provincial/urban/rural values. For educational attainment and dependency structure, we used the change rate derived from provincial projections and the historical urban and rural values in 2019 to generate projections at the urban and rural level.

Projections of disposable income, income inequality and income distribution. In this module, we first projected disposable income share of GDP (Y_1) and Gini coefficients (Y_2/PD_2) at provincial level and the income ratio (Y_3) and Gini ratio (Y_4) between urban and rural populations from 2020 to 2100 under the four future scenarios. Using these projections, we then solved for the future per capita disposable income at the provincial (PD_1), urban (SD_{1-1}), and rural level (SD_{1-2}), and urban and rural Gini coefficients ($SD_{2-1, 2-2}$). The provincial/urban/rural income distributions ($PD_3, SD_{3-1, 3-2}$) for the 31 Chinese provinces were then projected based on future Gini coefficients and per capita GDP.

The recursive projection approach. We developed an approach using recursive projections to create annual data of Y_1 – Y_4 from 2020 to 2100. In this approach, the RF model was trained on the most recent four years of data and then used to predict the response factors for each projected year. This process was repeated recursively from 2017 to 2099 to make projections for the years 2020 to 2100.

Solving the equality constraints. Based on the projected Y_1 and Y_3 , available provincial projections of Chinese GDP⁴², and urbanization rate⁴³, a system of linear equations was created and solved to generate the per capita disposable income at the provincial (PD_1), urban (SD_{1-1}), and rural (SD_{1-2}) level, as shown in Eq. 5. Based on the projected provincial Gini coefficient (PD_2), Y_4 , solved PD_1 and $SD_{1-1, 1-2}$, and future urbanization rate⁴³, the projections of urban (SD_{2-1}) and rural (SD_{2-2}) Gini coefficients were solved annually with equations shown in Eq. 6.

$$\begin{cases} Y_1 = \frac{I}{GDP} \\ Y_3 = \frac{I_{ur}}{I_{ru}} \\ I = PS_{ur} \times I_{ur} + PS_{ru} \times I_{ru} \end{cases} \quad (5)$$

$$\begin{cases} Y_4 = \frac{UrGini}{RuGini} \\ ProGini = PS_{ur}^2 \times \frac{I_{ur}}{I} \times UrGini + PS_{ru}^2 \times \frac{I_{ru}}{I} \times RuGini \\ \quad + PS_{ur} \times PS_{ru} \times \frac{I_{ur} - I_{ru}}{I} \end{cases} \quad (6)$$

Where PS_{ur} and PS_{ru} represent the urban and rural population share, respectively, while I_{ur} , I_{ru} , and I are the per capita disposable income at urban, rural, and provincial level.

Projections of income distribution. We assumed a log-normal distribution as the functional form of income distribution at the provincial, urban, and rural level. This is one of the most commonly assumed forms used in previous literature^{18,44}. We parameterized these using the projections of per capita disposable income and

Income distribution dataset of Chinese provinces at provincial, urban, and rural level

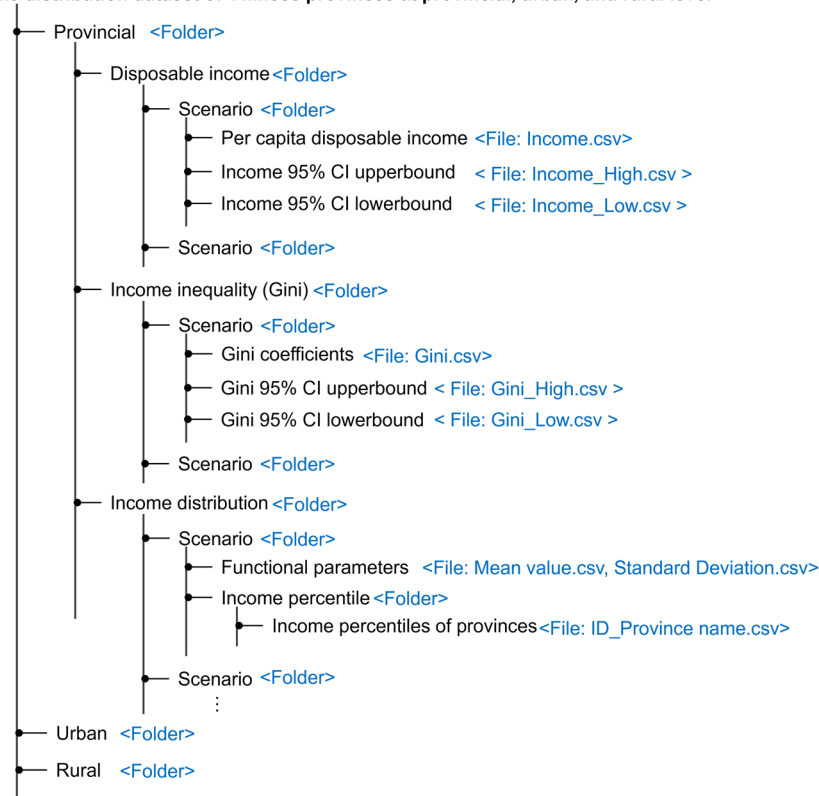


Fig. 3 Data organization. Dataset is available in the form of csv files.

Gini coefficients. Equation 7a–c describe the parameterization of the log-normal functional form¹⁶, applying a density distribution, which was defined and used for computing the income level at further different percentiles.

$$\sigma = 2 \times \operatorname{erf}^{-1}(Gini) \quad (7a)$$

$$\mu = \operatorname{Ln}(I) - \frac{\sigma^2}{2} \quad (7b)$$

$$F_x(x) = \varphi((\operatorname{Ln}(x) - \mu)/\sigma) \quad (7c)$$

Where *Gini* represents the Gini coefficients at provincial/urban/rural level for province *i* in year *t*, and *I* is the respective per capita disposable income.

Data Records

The projected yearly per capita disposable income, Gini coefficients, and income distribution (includes functional parameters and income percentile), under the four localized developmental scenarios are provided at the provincial, urban, and rural levels. These are all available in the public repository Figshare⁴⁵. This dataset also includes the 95% confidence intervals (CIs) of per capita disposable income and Gini coefficients for uncertainty analysis purposes. The dataset is available in the form of csv files, and Fig. 3 shows the hierarchy of data organization and file name templates.

To store the data, we define three main folders, named “Provincial”, “Urban”, and “Rural”, pertaining to the different spatial levels. Each main folder includes three sub-folders, named “Disposable income”, “Income inequality (Gini)”, and “Income distribution”. Each sub-folder contains four new folders named after the four scenarios to store the corresponding projections under different scenarios. In the scenario folders located in folder “Disposable income”, files named “Income.csv”, “Income_High.csv”, and “Income_Low.csv” are built to store per capita disposable income data (with unit of Yuan). For the scenario folders within the sub-folder “Income inequality (Gini)”, Gini coefficients and its 95% CIs under each scenario are stored in files “Gini.csv”, “Gini_High.csv”, and “Gini_Low.csv”, respectively. In scenario folders of sub-folder “Income distribution”, the parameters of income distribution stored in files “Mean value.csv” and “Standard deviation.csv”. Then, sub-folders “Income percentile” are further created within each scenario folder to store the files of income percentiles. All the files of income percentiles (with unit of Yuan) are named as “ID_Province name.csv”, while ID is the number assigned to each province.

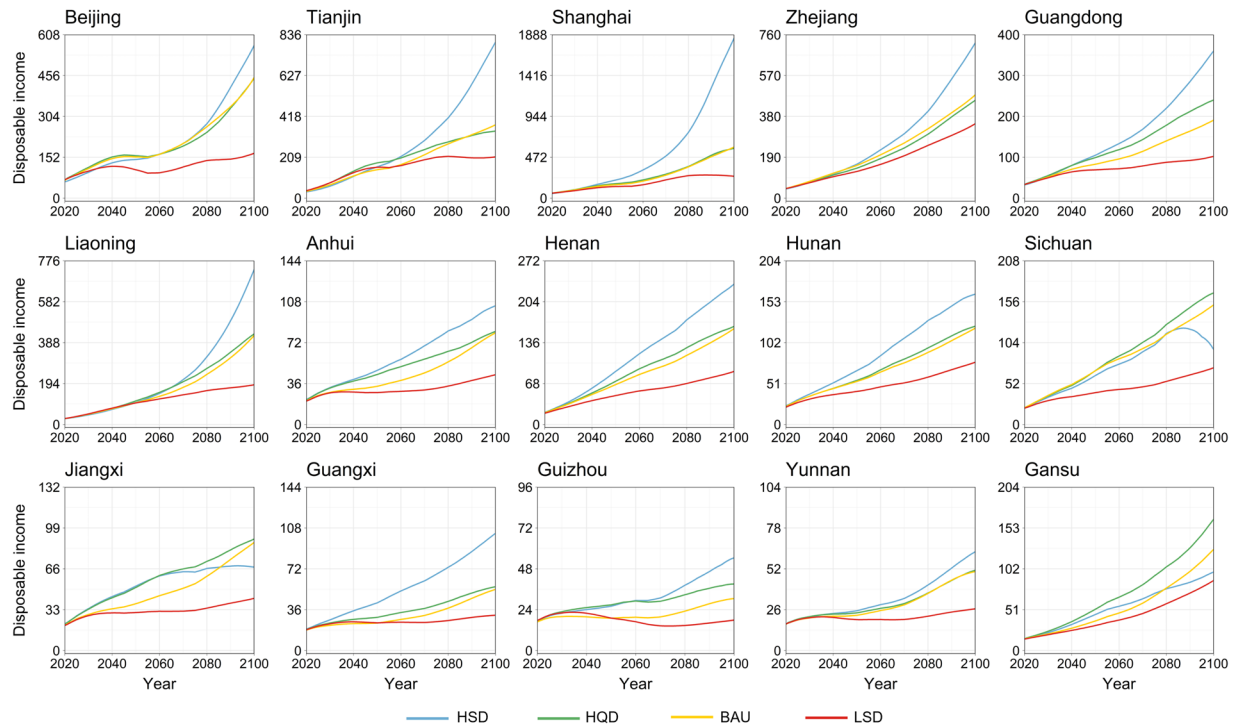


Fig. 4 Provincial per capita disposable income (thousand yuan) of sample provinces.

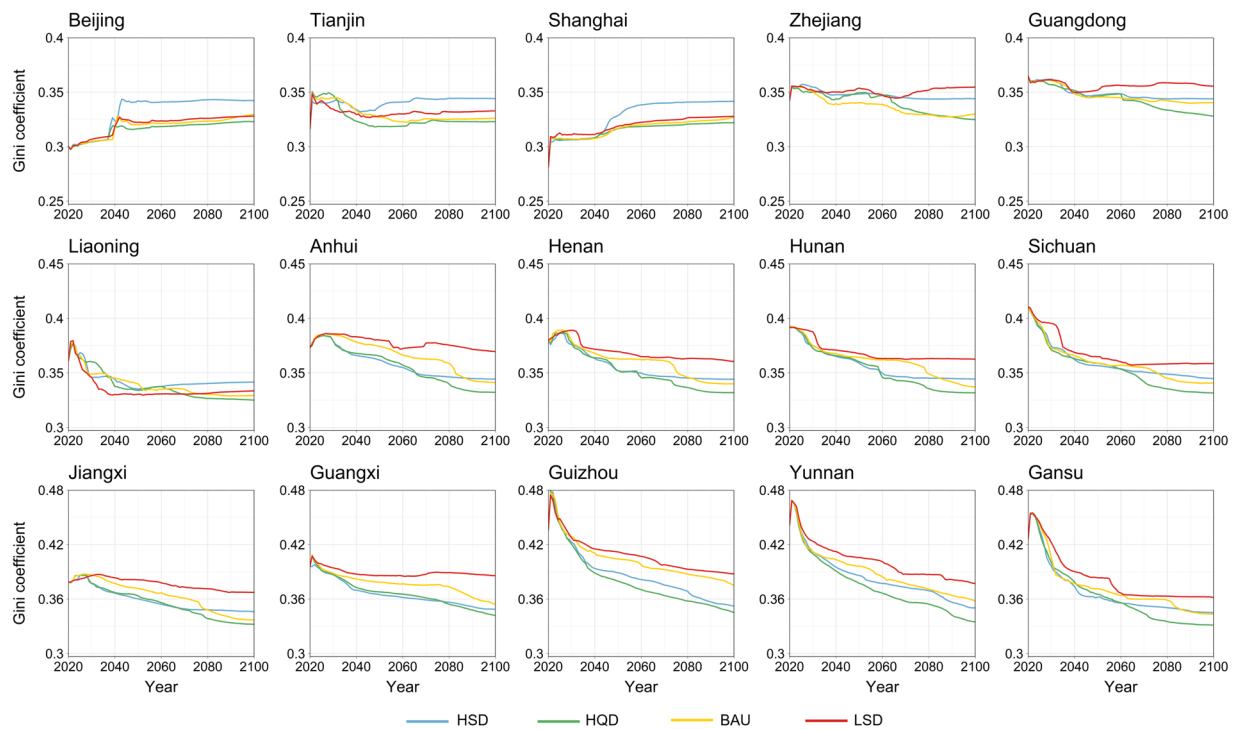


Fig. 5 Provincial income inequality (Gini coefficient) of sample provinces.

The provincial projections of per capita disposable income, Gini coefficients, and income distributions are shown in Figs. 4, 5. We distinguish the 31 provinces by three groups named tiers 1–3, based on their per capita GDP for the period 2007–2019. We then select five provinces from each tier to illustrate future disposable income and income inequality projections.

Indicators (%)	Models	Baseline validation			Robust validation (for Gini)		
		NL	FL	FLO	NL	FL	FLO
RMSE	Income	1.1	1.1	1.1			
MAPE		3.2	3.0	3.2			
RMSE	Gini	3.5	3.4	3.4	1.8	1.7	1.7
MAPE		5.4	5.3	5.4	3.9	3.7	3.5

Table 4. Model performance at provincial level.

Indicators (%)	Models	Baseline validation			Robust validation (for Gini)		
		NL	FL	FLO	NL	FL	FLO
RMSE	Income	2.0	2.2	2.1			
MAPE		1.9	2.0	1.9			
RMSE	Gini	4.8	5.0	4.4	14.1	14.7	15.3
MAPE		10.3	10.1	9.8	20.4	19.0	19.8

Table 5. Model performance at urban (rural) level.

Technical Validation

We tested the reliability and robustness of our results in the following steps, including model performance evaluation, errors assessment for provincial disposable income, and volatility analysis for provincial Gini coefficients.

Model performance evaluation. In Tables 4, 5, we describe the predictive capacities of the provincial and urban (rural) models. For provincial model, models trained on income dataset and Gini coefficient dataset all showed outstanding performance in both baseline validation and robust validation (for only Gini coefficient model). The RMSE of the models were all below 4%, and the MAPE was all below 6%, indicating that the RF models showed excellent predictive capacity of the temporal dimension and generalization ability in terms of the spatial dimension. For further analysis, we selected the model which exhibited the best performance, i.e., the income model using the FL dataset and the Gini model using FLO dataset to perform the subsequent procedures.

The urban (rural) level models for disposable income also showed satisfactory predictive performance across, particularly the model trained using the NL dataset, which produced a RMSE below 3% and a MAPE below 2%. However, the models trained on Gini coefficient ratio did not perform as expected. While the model trained on the FLO dataset had an acceptable performance for baseline validation, with a RMSE below 5% and a MAPE below 10%, it still did not meet expectations in the robust validation. This could be due to the uneven distributions of income equality across provinces, particularly in rural China⁴⁶, which suggests that the predictive variables used in this study to build the models were limited in capturing the spatial differences in rural Gini coefficients. In literature, several variables have been highlighted as important for explaining changes in rural Gini coefficients, such as employment rates across different industries⁴⁷, migration⁴⁸, and land use change⁴⁹. Nevertheless, data on these variables are rarely available, and their future projections carry considerable uncertainties. Therefore, we still regard the current model using the FLO dataset as the best choice for further simulation.

Error assessment for provincial disposable income. Table 6 presents the mean predictive errors from 2020 to 2023 between the provincial projections of per capital disposable income derived from per capita GDP and the provincial per capita disposable income collected from China Statistical yearbooks. The absolute percentage error (APE) is calculated based on Eq. 8 to reflect the predictive errors, where P_t represents the projected result and A_t represents the corresponding actual value.

$$APE (\%) = \left| \frac{P_t - A_t}{A_t} \right| \times 100\% \quad (8)$$

The mean APE across all 31 provinces is 4%, indicating a slight difference between projected income and actual value. Specifically, 29 among 31 provinces demonstrate APEs below 10%, and 24 among those 29 provinces show APEs below 5%.

Volatility analysis for income inequality projections. We cannot directly compare our projections with others' estimations due to the lack of similar income inequality datasets. To validate the reasonability and confidence of this dataset, we performed a volatility comparison based on provincial Gini coefficients projections. The volatility index was represented by the ratio of extreme deviation to minimum value.

To validate the ability of our model in predicting potential fluctuations in income inequality, we performed a volatility comparison between the projected provincial Gini coefficient and the historical Gini coefficient of a few countries, including the G7 countries and China. The Gini coefficients for these countries were obtained from the World Bank (<https://databank.worldbank.org/source/world-development-indicators>). The comparison results are shown in Fig. 6. The volatility of Gini coefficients across eight countries ranges from 8–45%, with

Province	APE (%)	Province	APE (%)
Beijing	3.08	Hubei	3.38
Tianjin	1.66	Hunan	2.34
Hebei	2.28	Guangdong	3.87
Shanxi	15.92	Guangxi	1.25
NeiMongol	11.90	Hainan	4.94
Liaoning	3.08	Chongqing	2.35
Jilin	2.05	Sichuan	2.03
Heilongjiang	1.58	Guizhou	3.94
Shanghai	1.15	Yunnan	3.32
Jiangsu	5.49	Xizang	9.39
Zhejiang	3.15	Shaanxi	2.51
Anhui	1.95	Gansu	4.37
Fujian	1.94	Qinghai	1.94
Jiangxi	2.65	Ningxia	5.23
Shandong	1.94	Xinjiang	7.26
Henan	6.85		

Table 6. The mean predictive errors from 2020 to 2023.

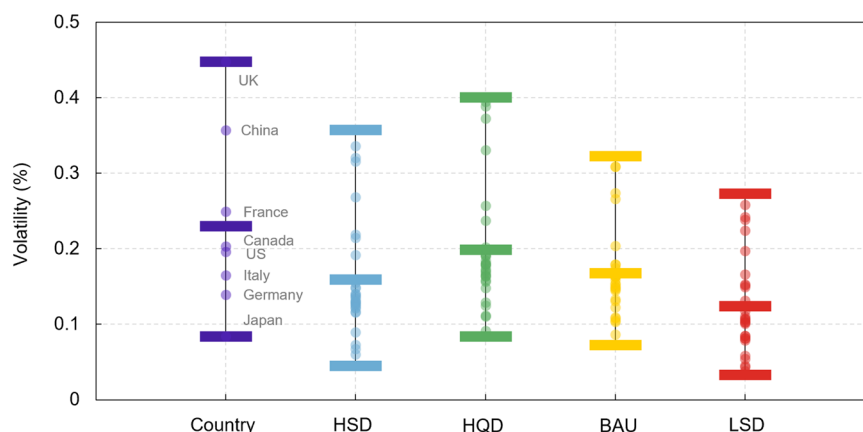


Fig. 6 The volatility comparison between Chinese provinces and selected developed countries.

an average of 23%. The volatility across provinces is 5–36% (16% average) for HSD, 8–40% (20%) for HQD, 7–32% (17%) for BAU, and 3–27% (12%) for LSD. Thus, the volatility range, we observe across provinces is of the similar range as that across countries and covers the volatility seen in the past few decades (20–60 years) of most countries. This indicates this dataset can capture potential fluctuations in income inequality on a long-term temporal scale.

Usage Notes

This study builds a methodological framework applying machine learning algorithms to project income inequality and distribution at the provincial, urban, and rural levels for 31 mainland Chinese provinces from 2020 to 2100 under different development pathways. In what follows, we discuss the potential applications of the proposed methodology and the released dataset, and we also interpret the uncertainties and limitations of this work.

Applicability of the methodology and dataset. Our products have several channels to easily interface with users' customized demands, and some examples of such uses are shown in Fig. 7. The first strand of applications for our products is to produce datasets that caters to users' customized demands. For example, this study provides a methodological framework to project income distribution datasets at different spatial level while considering necessary consistency constraints, which can be replicated and applied easily, as there are no strict limits on the form of data input (balanced or unbalanced panel data). Applying our methodology, users can produce similar datasets for other countries or regions at different spatial levels using their own historical datasets and assumptions regarding future scenarios.

In addition to such direct applications, the dataset produced by this study can also serve as an input for various research domains and analyses. For instance, the income distribution can be used to carry out further micro-level simulation-based analysis at the unit of individuals or households, so as to support highly granular

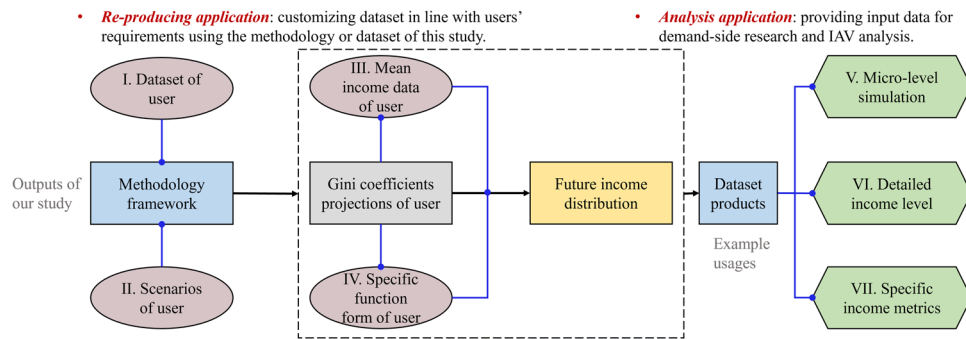


Fig. 7 The potential applications of the proposed methodology and released dataset.

analyses^{1,50}. Users can also determine specific income metrics according to their customized requirements, such as considering alternative international and national poverty thresholds, and diverse inequality metrics like the Palma ratio, or detailed income projections for all deciles. These data can serve as key input for various macro-level analyses across social, economic, and environmental domains, such as demand assessments, carbon footprint evaluations, and inequality research related to social well-being and health impacts of natural hazards. Meanwhile, this dataset can also be used in integrated assessment and computable general equilibrium models to clarify the coupled feedback between income, climate change, and economic outputs.

Uncertainties and limitations. We designed four different provincial-level pathways to explore divergent assumptions regarding future developments and related uncertainties in underlying socio-economic and demographic predictive factors. However, uncertainties still exist due to the lack of consideration of explicit policy interventions on income redistribution. While we consider indicators, such as educational expenditure, health expenditure, and social protection expenditure, as proxies for redistributive policies, we project these based on future development conditions rather than any explicit government intentions or policies to reduce inequalities. Therefore, the dataset released in this study can be regarded as a baseline reference range of disposable income, income inequality, and income distribution without taking possible policy intervention measures into account. In addition, our dataset was generated assuming a continuous development trend under each pathway, and thus does not include unforeseeable contingencies. Our dataset can be regarded as a benchmark and basis for further research that explores the impacts of specific policies, technological innovations, or events.

Code availability

All R codes for creating income inequality and distribution datasets for China provinces are stored in the public repository Figshare⁴⁵.

Received: 2 August 2024; Accepted: 13 December 2024;

Published online: 27 December 2024

References

- Poblete-Cazenave, M., Pachauri, S., Byers, E., Mastrucci, A. & van Ruijven, B. Global scenarios of household access to modern energy services under climate mitigation policy. *Nature Energy* **6**, 824–833 (2021).
- Zheng, L. *et al.* Health burden from food systems is highly unequal across income groups. *Nature Food* **5**, 251–261 (2024).
- Bruckner, B., Hubacek, K., Shan, Y., Zhong, H. & Feng, K. Impacts of poverty alleviation on national and global carbon emissions. *Nature Sustainability* **5**, 311–320 (2022).
- Mi, Z. *et al.* Economic development and converging household carbon footprints in China. *Nature Sustainability* **3**, 529–537 (2020).
- Asensio, O. I., Churkina, O., Rafter, B. D. & O'Hare, K. E. Housing policies and energy efficiency spillovers in low and moderate income communities. *Nature Sustainability* (2024).
- Wang, Q. *et al.* Examining energy inequality under the rapid residential energy transition in China through household surveys. *Nature Energy* **8**, 251–263 (2023).
- Hallegatte, S. & Rozenberg, J. Climate change through a poverty lens. *Nature Climate Change* **7**, 250–256 (2017).
- Sun, Y. *et al.* Global supply chains amplify economic costs of future extreme heat risk. *Nature* **627**, 797–804 (2024).
- Creutzig, F. *et al.* Towards demand-side solutions for mitigating climate change. *Nature Climate Change* **8**, 260–263 (2018).
- van Vuuren, D. P. *et al.* Alternative pathways to the 1.5°C target reduce the need for negative emission technologies. *Nature Climate Change* **8**, 391–397 (2018).
- Zimm, C. *et al.* Justice considerations in climate research. *Nature Climate Change* **14**, 22–30 (2024).
- Rao, N. D., van Ruijven, B. J., Riahi, K. & Bosetti, V. Improving poverty and inequality modelling in climate research. *Nature Climate Change* **7**, 857–862 (2017).
- Jing, C. *et al.* Gridded value-added of primary, secondary and tertiary industries in China under Shared Socioeconomic Pathways. *Scientific Data* **9** (2022).
- Rao, N. D., Sauer, P., Gidden, M. & Riahi, K. Income inequality projections for the Shared Socioeconomic Pathways (SSPs). *Futures* **105**, 27–39 (2019).
- Moyer, J. D. *et al.* How many people will live in poverty because of climate change? A macro-level projection analysis to 2070. *Climatic Change* **176**, 137 (2023).
- Narayan, K. B., O'Neill, B. C., Waldhoff, S. T. & Tebaldi, C. Non-parametric projections of national income distribution consistent with the Shared Socioeconomic Pathways. *Environmental Research Letters* **18** (2023).
- Fujimori, S., Hasegawa, T. & Oshiro, K. An assessment of the potential of using carbon tax revenue to tackle poverty. *Environmental Research Letters* **15** (2020).

18. Soergel, B. *et al.* Combining ambitious climate policies with efforts to eradicate poverty. *Nature Communications* **12**, 2342 (2021).
19. Casper, K. C. *et al.* Non-parametric projections of the net-income distribution for all U.S. states for the Shared Socioeconomic Pathways. *Environmental Research Letters* **18** (2023).
20. Hellebrandt, T. & Mauro, P. The future of worldwide income distribution. *Peterson Institute for international economics working paper* (2015).
21. van Ruijven, B. J. *et al.* Enhancing the relevance of Shared Socioeconomic Pathways for climate change impacts, adaptation and vulnerability research. *Climatic Change* **122**, 481–494 (2013).
22. Charpentier, A., Flachaire, E. & Ly, A. Econometrics and machine learning. *Economie et Statistique* **505**, 147–169 (2018).
23. Li, M. *et al.* Spatiotemporal dynamics of global population and heat exposure (2020–2100): based on improved SSP-consistent population projections. *Environmental Research Letters* **17** (2022).
24. Thorve, S. *et al.* High resolution synthetic residential energy use profiles for the United States. *Sci Data* **10**, 76 (2023).
25. Ahmed Gassar, A. A., Yun, G. Y. & Kim, S. Data-driven approach to prediction of residential energy consumption at urban scales in London. *Energy* **187** (2019).
26. Poblete Cazenave, M. & Pachauri, S. Household Energy Burdens in Europe following the Russian Incursion into Ukraine. (2023).
27. Li, Z., Wang, C. & Liu, Y. A dataset on energy efficiency grade of white goods in mainland China at regional and household levels. *Scientific Data* **10**, 445 (2023).
28. Sundrum, R. M. *Income distribution in less development countries.* (London and New York: Routledge, 1990).
29. Deaton, A. *The analysis of household surveys: a microeconomic approach to development policy.* (World Bank Publications, 1997).
30. Chen, D. & Ma, Y. Effect of industrial structure on urban–rural income inequality in China. *China Agricultural Economic Review* **14**, 547–566 (2022).
31. Zhou, Q. & Li, Z. The impact of industrial structure upgrades on the urban–rural income gap: An empirical study based on China's provincial panel data. *Growth and Change* **52**, 1761–1782 (2021).
32. Sauer, P., Rao, N. D. & Pachauri, S. in *Mobility and Inequality Trends* Vol. **30** 1–47 (Emerald Publishing Limited, 2023).
33. Gradin, C. in *Income inequality around the world* Vol. **44** 109–177 (Emerald Group Publishing Limited, 2016).
34. Kakwani, N., Neri, M. C. & Son, H. H. Linkages between pro-poor growth, social programs and labor market: the recent Brazilian experience. *World Development* **38**, 881–894 (2010).
35. Muinelo-Gallo, L. & Roca-Sagalés, O. Joint determinants of fiscal policy, income inequality and economic growth. *Economic Modelling* **30**, 814–824 (2013).
36. Jianu, I. The impact of government health and education expenditure on income inequality in European Union. *Theoretical & Applied Economics*, (2018).
37. Chevan, A. & Stokes, R. Growth in family income inequality, 1970–1990: Industrial restructuring and demographic change. *Demography* **37**, 365–380 (2000).
38. Maia, A. G. & Sakamoto, C. S. The impacts of rapid demographic transition on family structure and income inequality in Brazil, 1981–2011. *Population studies* **70**, 293–309 (2016).
39. Chen, Y. *et al.* Provincial and gridded population projection for China under shared socioeconomic pathways from 2010 to 2100. *Scientific Data* **7**, 83 (2020).
40. Han, X., Wei, C. & Cao, G.-Y. Aging, generational shifts, and energy consumption in urban China. *Proceedings of the National Academy of Sciences* **119**, e2210853119 (2022).
41. Mason, A. & Racelis, R. A comparison of four methods for projecting households. *International Journal of Forecasting* **8**, 509–527 (1992).
42. Jing, C. *et al.* A gridded dataset comprising value-added of primary, secondary and tertiary industries in China under shared socioeconomic pathways from 2020–2100. Version 2. *4TU.ResearchData*. dataset. <https://doi.org/10.4121/14113706.v2> (2022).
43. Chen, Y. *et al.* Provincial and gridded population projection for China under shared socioeconomic pathways from 2010 to 2100. *figshare* <https://doi.org/10.6084/m9.figshare.c.4605713.v1> (2020).
44. Riahi, K. *et al.* The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global environmental change* **42**, 153–168 (2017).
45. Lei, M., Pelz, S., Pachauri, S. & Cai, W. A dataset of income distribution on provincial, urban, and rural levels for China from 2020 to 2100. *figshare* <https://doi.org/10.6084/m9.figshare.27888801> (2024).
46. Gao, J., Liu, Y., Chen, J. & Cai, Y. Demystifying the geography of income inequality in rural China: A transitional framework. *Journal of Rural Studies* **93**, 398–407 (2022).
47. Butler, J., Wildermuth, G. A., Thiede, B. C. & Brown, D. L. Population Change and Income Inequality in Rural America. *Population Research and Policy Review* **39**, 889–911 (2020).
48. Howell, A. Impacts of Migration and Remittances on Ethnic Income Inequality in Rural China. *World Development* **94**, 200–211 (2017).
49. Bou Dib, J., Alamsyah, Z. & Qaim, M. Land-use change and income inequality in rural Indonesia. *Forest Policy and Economics* **94**, 55–66 (2018).
50. Pachauri, S., Poblete-Cazenave, M., Aktas, A. & Gidden, M. J. Access to clean cooking services in energy and emission scenarios after COVID-19. *Nature Energy* **6**, 1067–1076 (2021).
51. Calzadilla, A. Global income distribution and poverty: implications from the IPCC SRES scenarios. (Kiel Working Paper, 2010).

Acknowledgements

This work was supported by the National Key R&D Program of China (2023YFF0805901); the National Natural Science Foundation of China (72091514); China Meteorological Administration Climate Change Special Program (CMA-CCSP); the Youth Innovation Team of China Meteorological Administration (CMA2023QN15); Tsinghua-Rio Tinto Joint Research Center for Resource Energy and Sustainable Development, Tsinghua University.

Author contributions

W.J. Cai and M.Y. Lei conceived this study. M.Y. Lei collected the data and performed data cleaning. M.Y. Lei, S. Pelz, and S. Pachauri developed methodology framework and produced this dataset. M.Y. Lei wrote the initial draft of the manuscript. All authors discussed the results and edited the manuscript.

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/s41597-024-04304-x>.

Correspondence and requests for materials should be addressed to W.C.

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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2024