ESSD: Response to reviewers for Narayan et al. ("A consistent dataset for the net income distribution for 184 countries, aggregated to 32 geographical regions and the world from 1958-2015")

Summary of content-

<u>Reviewer 1</u>

General comments

"The paper attempts to create a large dataset of 190 countries over almost 70 years providing consistent information on net income distributions. Such an attempt is valuable. However, if such a database does not already exist, or with limited scope such as the LIS, it is because it raises serious challenges, ultimately related to the lack of suitable data. I found the paper unconvincing in the ways it tackles these challenges. I thus believe that the database it intends to produce (and document) is unlikely to be taken up by other researchers and institutions."

Response: Thank you for your detailed review of our manuscript. We have responded to each of your specific comments below. Here, we would like to mention that we have edited the manuscript to better explain the purpose of our dataset construction. In particular, we constructed this dataset to calibrate inequality metrics in regional and global integrated assessment models (IAMs). These models require income distribution data that is comparable across countries. As the reviewer correctly notes, there are surveys conducted in individual countries in individual years, however it is difficult to extract comparable metrics across countries from these datasets for income groups. The LIS and the PovCal are the only two surveys which produce metrics comparable across countries, hence we started with these sources. But even these sources contain a mix of income concepts which is why our imputation (between consumption and net income) was necessary. But, we have clarified that such an imputation is applied to only a small subset of observations in our dataset (394 out of 8522). Finally, models such as GCAM operate at regional scales, some of which aggregate across countries. Thus, we document national income distributions aggregated to the GCAM regional level. Thus far, most IAMs to our knowledge have used income inequality data from LIS and PovCal and have used the different definitions of the income concept interchangeably, hence our dataset is an improvement upon the existing literature. These points have been elaborated upon in our revised manuscript, especially in the introduction and discussion.

Specific Comments

1. *"Imputing net income shares using consumption shares:*

a. *I am skeptical of this approach without more information about the estimation sample and the country-years for which such imputation is performed. If the two sets of countries are different, there are reasons to doubt that good R-squares would translate into good out-of-sample predictions.*

b. *I understand that the current setup with 10 regressions does not ensure that all income shares add up to one. Why not run one regression and impose the constraint that the sum of all income shares must be equal to 1?"*

Response:

a. This has been addressed in section 2.3 of the paper in detail (We have now provided more detailed information here). In particular,

i.) we used a dataset of 257 country-year observations which had data for both net income and consumption.

ii.) We split this dataset into a training data set (all pre 2004 observations) and a testing dataset (observations starting 2004 onwards) and fit ten separate regressions where we impute individual net income deciles from consumption deciles. See below for the validation of our imputation approach.

iii.) Most of these regressions had an R squared of over 0.6 except the regression for d9 which was 0.29.

iv.) For this reason, we impute net income shares for 9 deciles (all deciles excluding d9) and then calculate d9 as the residual. This resulted in all deciles adding up to 1 for all country-years, which we have verified. We have now made this clear in the text.

v.) If the regression introduced inconsistencies between deciles (e.g., d7 > d8), the GINI coefficient thus calculated would yield an incorrect number. Therefore, we re-calculated a GINI coefficient from our imputed deciles to ensure that there are no inconsistencies between deciles.

<u>Validation of fit for our imputation method-</u> To validate our imputation method we calculated errors (Imputed shares - actual shares) for our testing dataset (n=123). We compared the error by decile for the dataset (See Figure below). This figure below is attached as SI 2 Figure 4. The mean error across deciles is generally within half a percent across all years. There are larger differences for the year 2011, where we had very few observations.

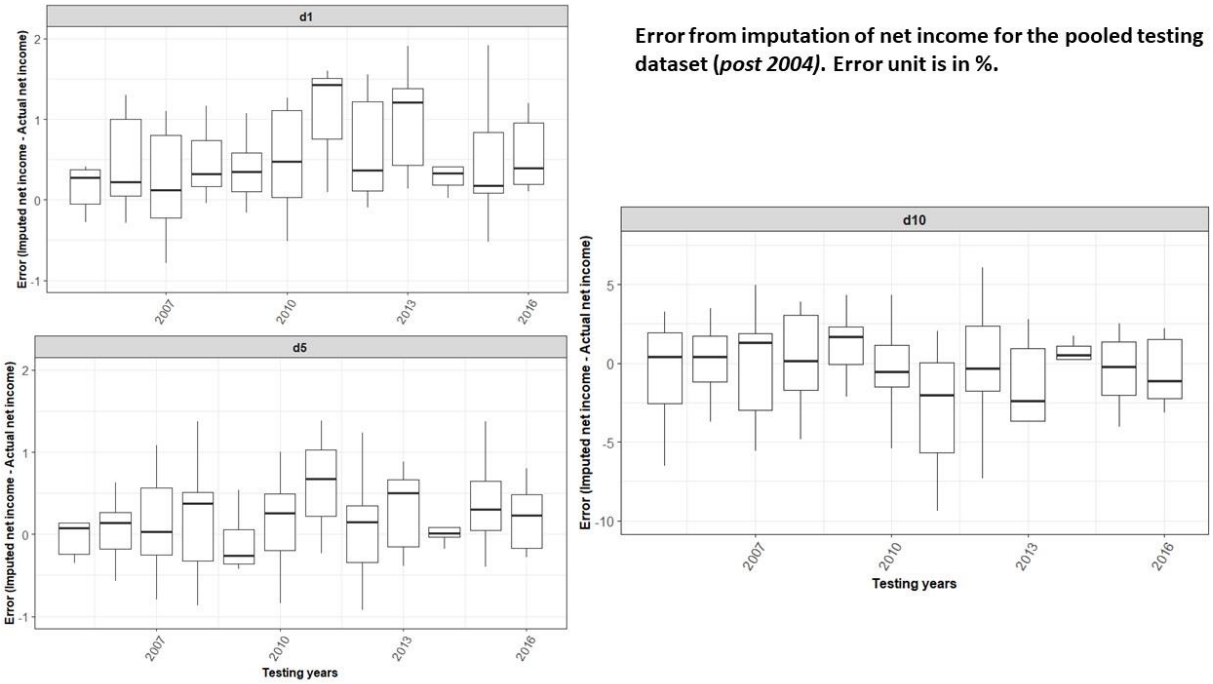

Figure : Percent Error (imputed income deciles- actual) for the testing dataset. Error is shown for 3 deciles, namely d1, d5 and d10 for all years in the testing dataset .

Similarly, we also compared the fit for individual countries from our testing dataset (Figure below). Once again we note that the fit is reasonable across deciles for individual countries. We are able to reproduce the R squared documented from the training dataset in our plot below. Also note that even when R squared values based on the testing data are lower, all imputed values are within a 95 percent confidence interval of actual values. Note that the figure below is attached as Figure 4 in the revised paper.

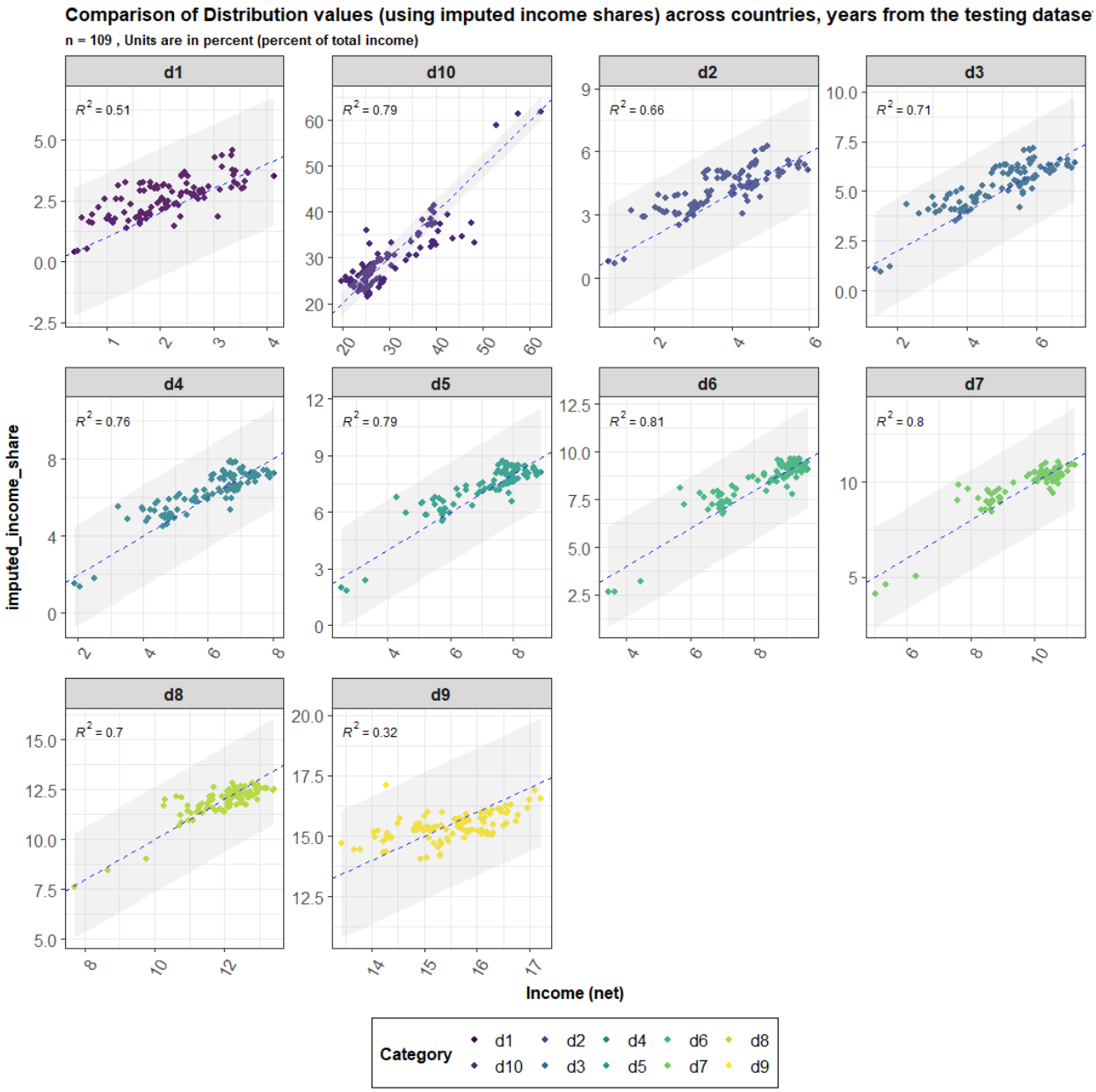

Figure: Comparison of imputed and actual net income decile values for the testing dataset across all deciles. We also show the R squared from the fit here.

b.) As noted in point iv.) and v.) above we have noted that our current approach produces deciles that add up to 1. We would also like to note that the imputation affects a small subset of data points (394 out of 8522). The majority of other observations are calculated using the PCA algorithm, whose fit has been clarified in more detail below.

2. *"Imputing net income deciles based on summary measures of the Gini coefficient*

a. *The same point mentioned above about in-sample vs. out-of-sample predictions applies here too.*

b. *Where it is not known whether the Gini coefficients are based on income or consumption, it would be best to drop these countries and years to ensure consistency of the income concept."*

Response:

a.) Thank you for this comment. We introduced the PCA algorithm in Narayan et al. 2023- https://iopscience.iop.org/article/10.1088/1748-9326/acbdb0/meta, where we extensively validated the fit of the algorithm (both in-sample and out-of-sample). We examined the fit for our pooled dataset when compared to other methods (See Figure 1 and Figure 4 of that paper) and we also produced comparisons for individual countries and deciles (See SI Figure 12 and SI Figure 13 of that paper). Our algorithm was found to provide a better fit across all deciles and countries. We have now added more text related to the PCA algorithm fit when applied to our dataset. We could bring in more figures from that paper if useful. We have attached below the main figure from that paper which shows the improved fit for the pooled dataset (The orange dots below represent the fit based on the PCA model)-

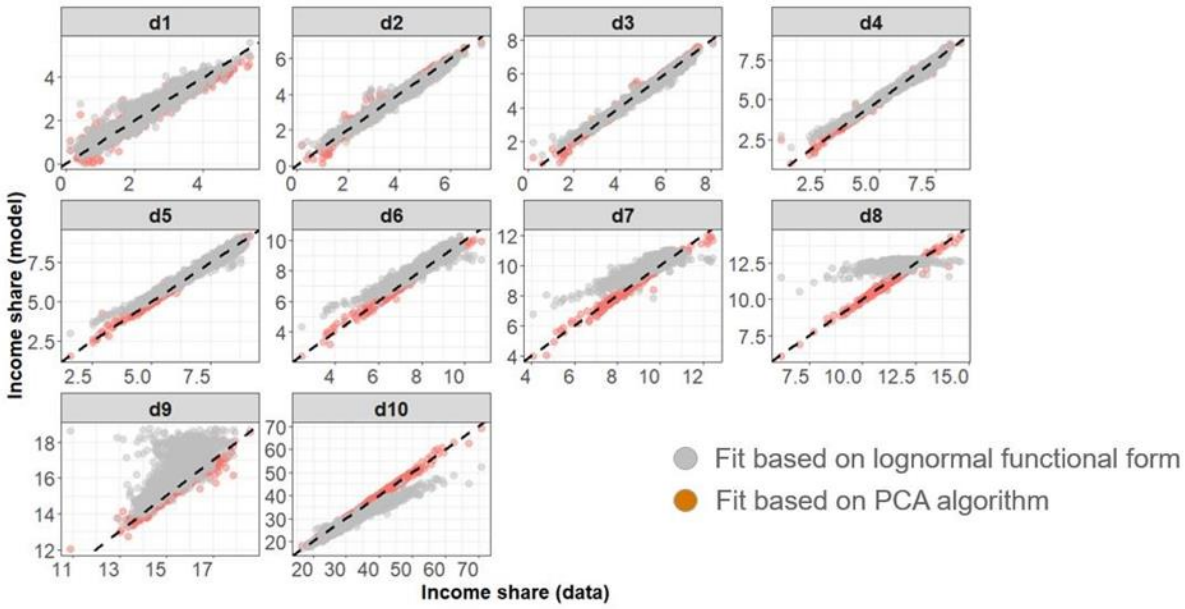

Figure : Comparison of fit of lognormal functional form (grey dots) with PCA based fit (orange dots) with data for each decile (facet). Lines represent 1 to 1 fit between x and y axis. Income shares are expressed as a percent of total income.

b.) Thank you for this very insightful comment. Regarding the concern about using observations when the GINI value is based on consumption data, as noted in section 2.4, we have now identified all observations imputed from a consumption GINI. These have been marked separately in the dataset itself. Users of our dataset have informed us that they would still need a full time series irrespective of the imputation method, hence these observations are still retained. We have updated the table 4 in the paper with the revised

statistics regarding the observations. We agree with the reviewer that this is an issue that deserves attention and could be revisited in the future.

Data source	count of observations
Original data on net income	1191
Imputed from original data on consumption	494
Imputed from Net income GINI coefficient	4201
Imputed from Expenditure & Consumption GINI coefficient	1303
Imputed from Gross income GINI coefficient	1333
Total	**8522**

Table : *Summary of observation types in final data set*

3. "Aggregating income distributions to the regional level
a.) I do not see a clear motivation for this section (other than the need for the authors to carry out these analyses for another project/report).
b.) The approach sounds highly problematic as it appears to confuse (or ignore the differences) between household net income and GDP per capita. In addition, it also appears to ignore (crucial) variations in income dispersion within income decile groups.
c.) The same issues apply to section 4, in which the authors aggregate country income distributions up to the global level."

Response:

a.) Thank you for this comment. We have clarified in more detail why this step was necessary. In particular, we constructed this dataset to calibrate inequality metrics in regional and global Integrated Assessment Models (IAMs) such as GCAM. Regional models such as GCAM require regional boundary conditions, thus we developed a method to aggregate national income distributions to GCAM regional levels as an example. If national data on income distribution was to be used in any other model , such an aggregation method would be necessary. This has been clarified in section 3 of the manuscript now.

b.) Thanks again for the comment. We agree that our method for aggregation is subject to uncertainties and limitations. Firstly, regional economic models use GDP per capita as a proxy for income levels, hence we used the same variable to perform the aggregation. We

can alternatively use net income; however, we wanted to be consistent with the model's measure of income. We have noted that our current method ignores within-decile variations in income levels and assumes a uniform distribution of income within a decile. This can and should be improved upon as more data on income distribution become available. We have clarified this in our manuscript.

c.) As noted in point b.) above, we agree that our method is subject to limitations. We agree with the reviewer that aggregation to the global level would introduce more uncertainties and have dropped this section from the manuscript. We have also edited the title of the paper.

4. "The differences between the different data sources shown in Figure 5 are concerning. Given these sizeable differences, it is far from clear that one could accurately assess inequality levels and trends using data imputed by the authors."

Response: The largest difference here is noted for the US. As noted in the previous comment, this is because we used the GINI data based on gross income to estimate inequality in the US when no data was available. This is because the ACS data, which produces data on income distributions in the US, is based on gross income as opposed to net income. We have dropped the observations based on the gross income GINI when constructing the figure. When we re-made the figure solely based on net income observations or those imputed from a net income GINI, this solved our issue as noted in figure 5, also attached below. Note that there can still be jumps between years (For example d10 income shares in India by 7% between 2005 and 2010) but this is possible since the survey design can change between years and between data sources themselves in different years. However, with our method, jumps between years are limited.

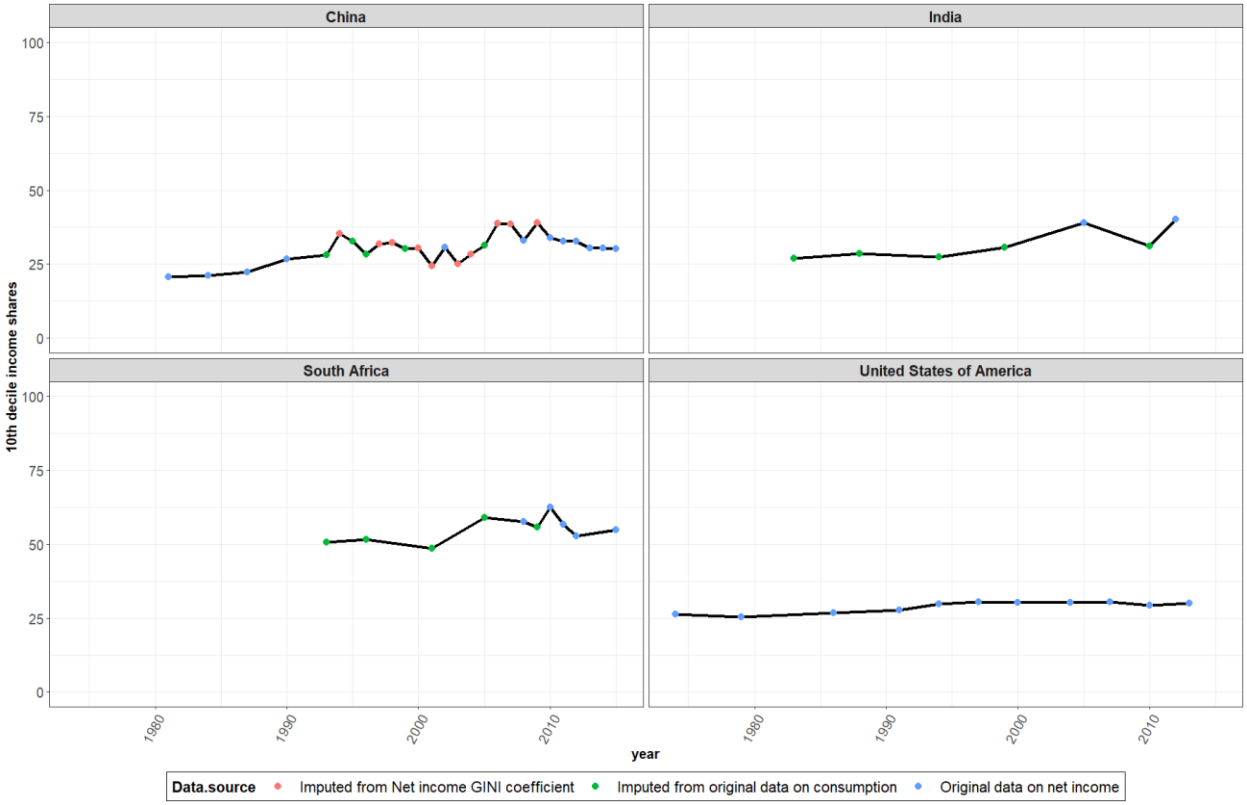

Figure : Temporal trends in the 10th decile for the complete dataset. Colors represent different data sources.

5. "*US results (Fig 5): why is "original data" for net income not available for the whole period? The CPS is a large and representative survey that collects detailed income information and that has been running yearly since the 1960s."*

Response: Thanks for this comment. As the reviewer correctly notes, there are surveys conducted in individual countries in individual years (such as the CPS), however it is difficult to extract comparable metrics across countries from these datasets for income groups. The LIS and the PovCal are the only two datasets which produce metrics comparable across countries, hence we started with these sources. The CPS is based on consumer expenditures, and its Annual Social and Economic Supplement (ASEC) records only gross income. Neither produce the income concept that we are interested in, namely net income. We have clarified this point throughout our manuscript.

6. "Introduction : the paragraph starting on line 29 is odd because it suggests that "at the national level", datasets on income inequality have been "limited to summary metrics". That is clearly not true. In many countries, detailed microdata allows researchers and statistical agencies to produce detailed distributional analyses."

Response: As the reviewer correctly notes, there are surveys conducted in individual countries in individual years which produce very useful microdata, however it is difficult to extract comparable metrics across countries from this microdata for income groups. The LIS and the PovCal are the only two datasets which produce metrics comparable across countries, hence we started with these sources. We agree that microdata is available across countries and have acknowledged that in our manuscript now.

Reviewer 2

1. *"The paper contributes to a growing literature on world databases of incomes. The data sources selected are appropriate but insufficient. There is no mention of the work led by Branko Milanovic in the US and Thomas Piketty in France. Both these authors oversee global data efforts to measure income and income inequality that should be reviewed/acknowledged by the authors."*

Response: Thank you for the careful review of our manuscript and your comments. As noted in the responses to reviewer one, we have added text in section 2.3 to address these points. We have acknowledged the Milanovic and Piketty approaches in our citations and text and agree that it is important to highlight such efforts in this space. However, we have noted that the Lanker-Milanovic dataset is still a combination of PovCal and LIS and has limited temporal coverage (the data is only available to the year 2013). The Piketty dataset is available only for the USA and is not global.

We have not drawn on these data directly because we constructed this dataset to calibrate inequality metrics in regional and global economic models. These models require income distribution data that is comparable across countries i.e. the same income concept.

2. *"The central contributions of the paper are the regression-based approach to estimate net incomes from consumption data and the PCA approach used to estimate income data from national Gini coefficients. The regression-based approach (section 2.3) is popular but there are better methods that should provide better fit including quantile regressions and random forest methods. The problem with simple OLS is that they are very poor at predicting incomes on the tails of a distribution (the predicted income distribution is always much narrower than the original income distribution). The fact that the authors run different regressions for each decile accentuates this problem by creating discontinuities between deciles. This problem can be overcome with quantile regressions or, better, with random forest. The probabilistic nature of random forest fits the tails if income distributions much better than standard OLS. I would recommend the authors to test both methods and compare results with the current ones."*

Response: Thank you for this very thoughtful comment. We agree that the linear regression we implemented for our imputation is simplistic. However, we justify its current usage based on several points,

a.) We first note that the imputation affects a small subset of data points (394 out of 8522). The majority of other observations are calculated using the PCA algorithm, whose fit has been clarified in more detail in the paper (as we also describe below).

b.) We have now described our approach for imputation in more detail and added new validation information, namely-

i.) we used a dataset of 257 country-year observations which had data for both net income and consumption.
ii.) We split this dataset into a training data set (all pre-2004 observations) and a testing dataset (observations from2004 onwards) and fit ten separate regressions where we impute individual net income deciles from consumption deciles.
iii.) Most of these regressions had an R squared of over 0.6 except the regression for d9 which was 0.29.
iv.) We impute net income shares for 9 deciles (all deciles excluding d9) and then calculate d9 as the residual. This resulted in all our imputed deciles adding up to 1.
v.) We re-calculated a GINI coefficient from our imputed deciles to ensure that there are no inconsistencies between deciles. Note that if the regressions introduced inconsistencies (e.g., if imputed d7 is higher than imputed d8), the GINI coefficient calculation would result in implausible values.
c.) We also performed several types of validation for our imputation-

To validate our imputation method, we calculated errors (Imputed shares- actual shares) for our testing dataset (n=123). We compared the error by decile for the dataset (See Figure 1 below). The mean error across deciles is generally close to zero across all years. There are larger differences for the year 2011, where we have very few observations.

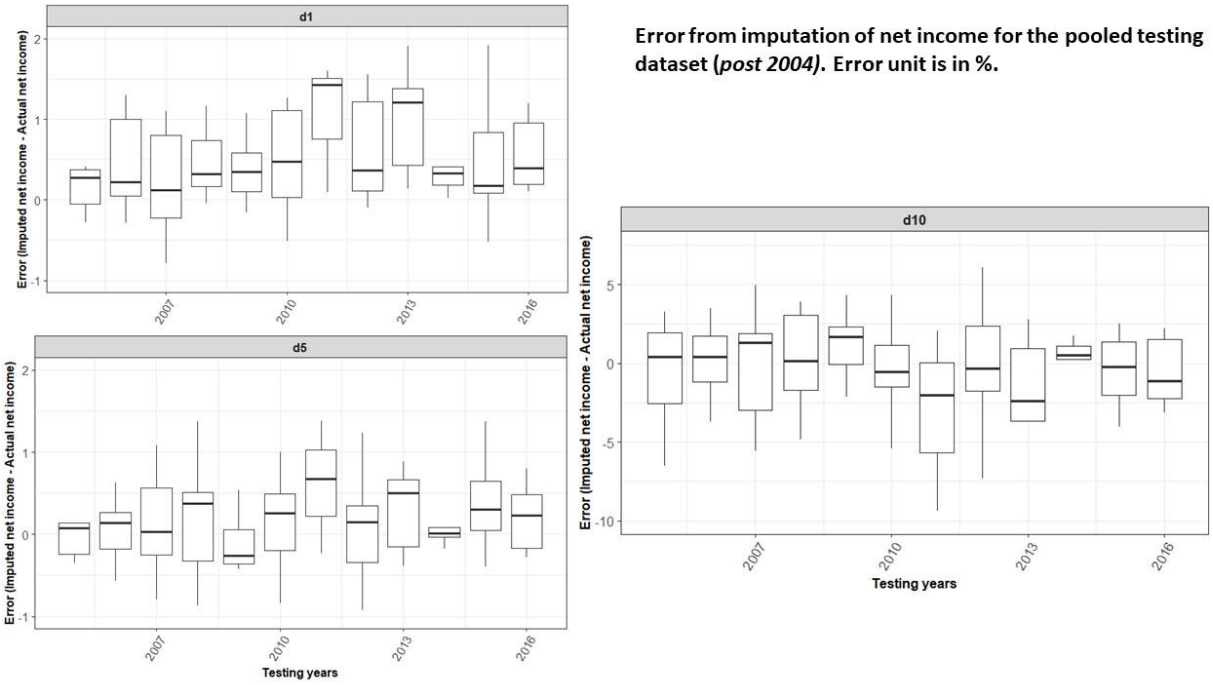

Figure : Percent Error (imputed income deciles- actual) for the testing dataset. Error is shown for 3 deciles, namely d1, d5 and d10 for all years in the testing dataset .

Similarly, we also compared the fit for individual countries from our testing dataset (Figure 2). Once again we note that the fit is reasonable across deciles for individual countries. The Figures shown here are attached in the revised paper as SI 2 Figure 4 and Figure 4 respectively.

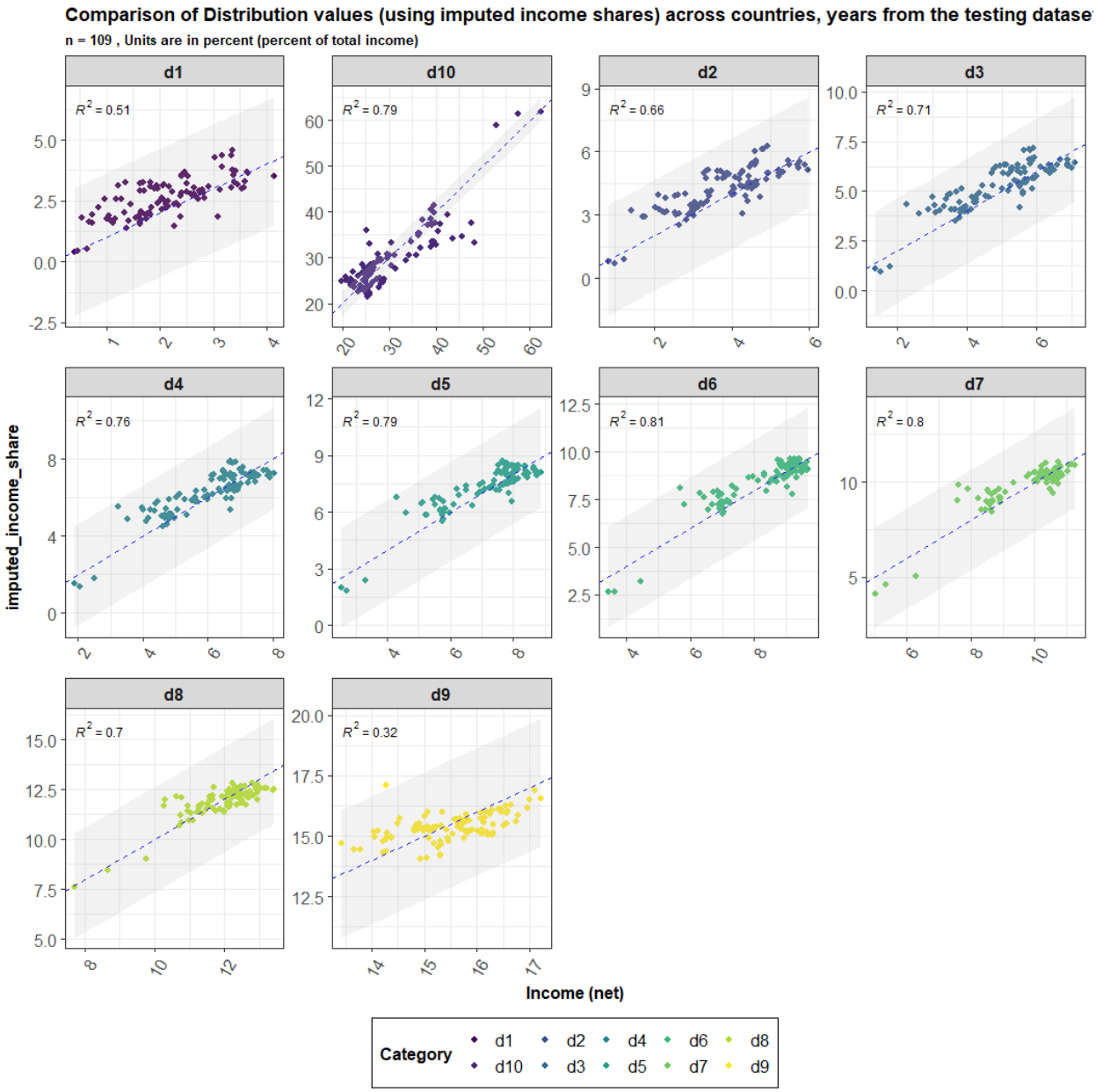

Figure: Comparison of imputed and actual net income decile values for the testing dataset across all deciles. We also show the R squared from the fit here.

3. *"The PCA method (section 3.4) is somewhat unconventional for this specific literature. This, per se, is not a critique, but it does require validation beyond what the authors offer. Here I would suggest taking the entire net income distribution for a few countries where these data are publicly available, calculate the net income deciles and the Gini coefficient, plug this Gini into equations 3) and 4) and compare the resulting estimated deciles with those calculated from the full data. Also, it is important to clarify where the coefficients in equations 3 and 4 come from. I could not find the model and the results of the "equation estimated on 1659 observations".*

Response: We introduced the PCA algorithm in Narayan et al. 2023- https://iopscience.iop.org/article/10.1088/1748-9326/acbdb0/meta, where we extensively validated the fit of the algorithm. We examined the fit for our pooled dataset when compared to other methods (See Figure 1 and Figure 4 in that paper) and we also produced comparisons for individual countries and deciles (See SI Figure 12 and SI Figure 13 in that paper). Our algorithm was found to provide a better fit across all deciles and countries (See Figure attached below). We have now added more text related to the PCA algorithm fit for our dataset in this paper. We can bring in more figures from our older paper here as well, but we leave that to the discretion of the editor.

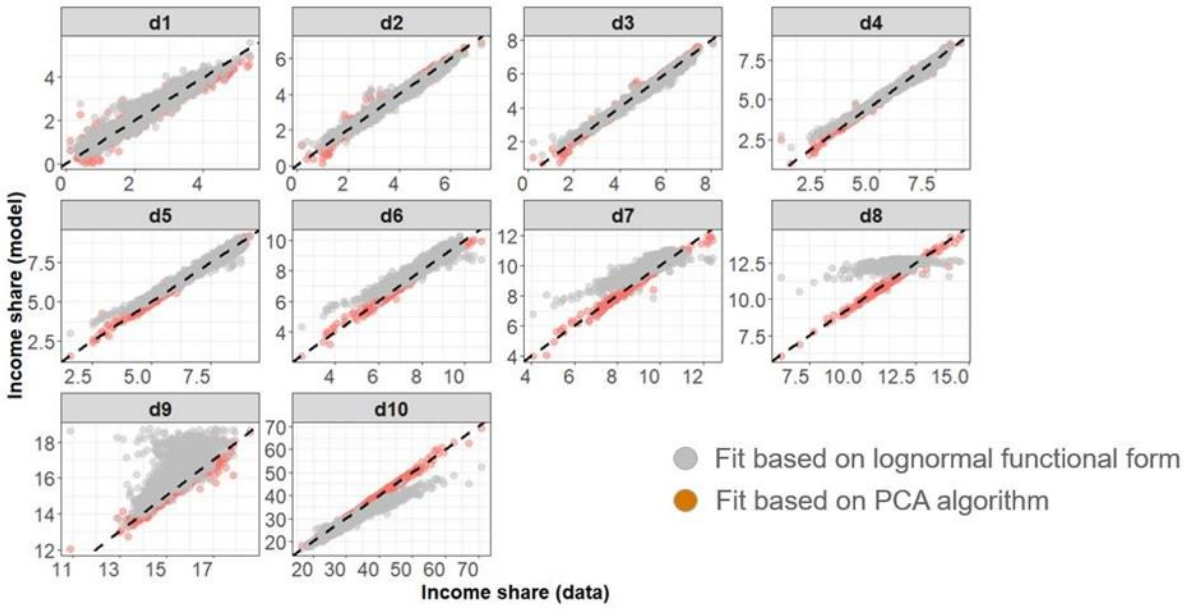

Figure : Comparison of fit of lognormal functional form (grey dots) with PCA based fit (orange dots) with data for each decile (facet). Lines represent 1 to 1 fit between x and y axis. Income shares are expressed as a percent of total income.

4. *"The revisions suggested above are substantial and I would recommend the authors to cut out of the paper the work on regions, which is important for the GCAM but a distraction from the main objective of the paper. Instead, the regional work could be the object of a separate paper. This strategy would also allow the authors to target different audiences better."*

Response: Thank you for this comment. We have clarified in more detail why this step was necessary. In particular, we constructed this dataset to calibrate inequality metrics in regional and global economic models such as GCAM. Regional models such as GCAM operate on regional boundary conditions and hence it was necessary to produce a method to aggregate national income distributions to the regional level. This ensures that models can be effectively calibrated. If national data on income distribution was to be used in any other model, such an aggregation method would be necessary. This has been clarified in

section 3 of the manuscript now. Also, we removed the section of the paper on the aggregation to the global level, as regional models do not need to do such an aggregation. We also responded in more detail to this point based on comments from Reviewer 1.