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")

(Revisions for Round 2)

Summary of content-

<u>Reviewer 1</u>

General comments

"The paper describes an empirical dataset on a large set of countries and time periods covered of household deciles, which are a very useful exercise.
Overall it is well written, and the produced data looks indeed very good and addresses several issues in existing data sets. However, I have a few concerns on the presentation and questions about few references that would be good to take into account before publication."

Response: Thank you for your detailed review of our manuscript. We have made several edits in response to your comments. In particular, we added more details with respect to data coverage, and have documented several points in the methodology section more clearly. The responses to the specific comments are attached below. Note that we have mentioned specific sections with page numbers, line numbers where edits are made.

Specific Comments

1. *"Data coverage: Table 6 is not very informative (average populations of countries not very useful, maybe total population/GDP covered or not would be more meaningful, best in a plot or map). And what are the biggest countries missing? Secondly, YEARs: Why does you r data series stop in 2015? In 2023 this is a big issue for using the data for empirical researchers, is there a reason? Finally, a heat map of all years and countries would be nice. Is it a panel without gaps ultimately?"*
   Response: Thank you for the insightful comments. In response to the same,

   i.) **We have edited Table 6 to represent the % of GDP and population covered. We utilized the GDP and population data from the SSP database for the same. The same is attached below. As seen in the table a majority of the global population and global GDP is covered in our dataset. We also added the % of countries that are imputed from the GINI (using the PCA algorithm) and the ones imputed from consumption (using the regression equation).**

Country data status	year	Global GDP|PPP	Global Population
Data not available	**2010**	**0.4%**	**2.0%**
Data not available	**2015**	**0.3%**	**1.3%**
Imputed from GINI coefficient (using PCA algorithm)	2010	19.9%	25.8%
Imputed from GINI coefficient (using PCA algorithm)	2015	45.1%	52.5%
Imputed from original data on consumption (Using regression)	2010	10.8%	31.2%
Imputed from original data on consumption(Using regression)	2015	5.8%	9.6%
Original data on net income	2010	68.9%	41.0%
Original data on net income	2015	48.8%	36.6%

Table 6: Coverage by data status in terms of GDP in PPP and Population from the SSP database V9.

ii.) **We have also noted the major countries that are missing in section 4 (Line 2-7 on Page 22 of the revised manuscript). Specifically, we note "** ***found that the countries that are missing data in the latest historical year (2015) only constitute 1.3% of the global population and 0.3% of the global GDP. The biggest countries that are missing data in terms of population in 2015 are Morocco (33 million people), North Korea (24 million people) and Somalia (10 million people). In terms of GDP, the biggest countries missing are Morocco (123 billion USD at PPP), Oman (68 billion USD at PPP) and Equatorial Guinea (18 billion USD at PPP)*****"**

iii.) **In addition to the edits to Table 6, we also show a map of countries by data status in Figure 11 (attached below). This allows the user to understand countries that are missing recent data on income distributions. We further added all categories in our final dataset to the map below.**

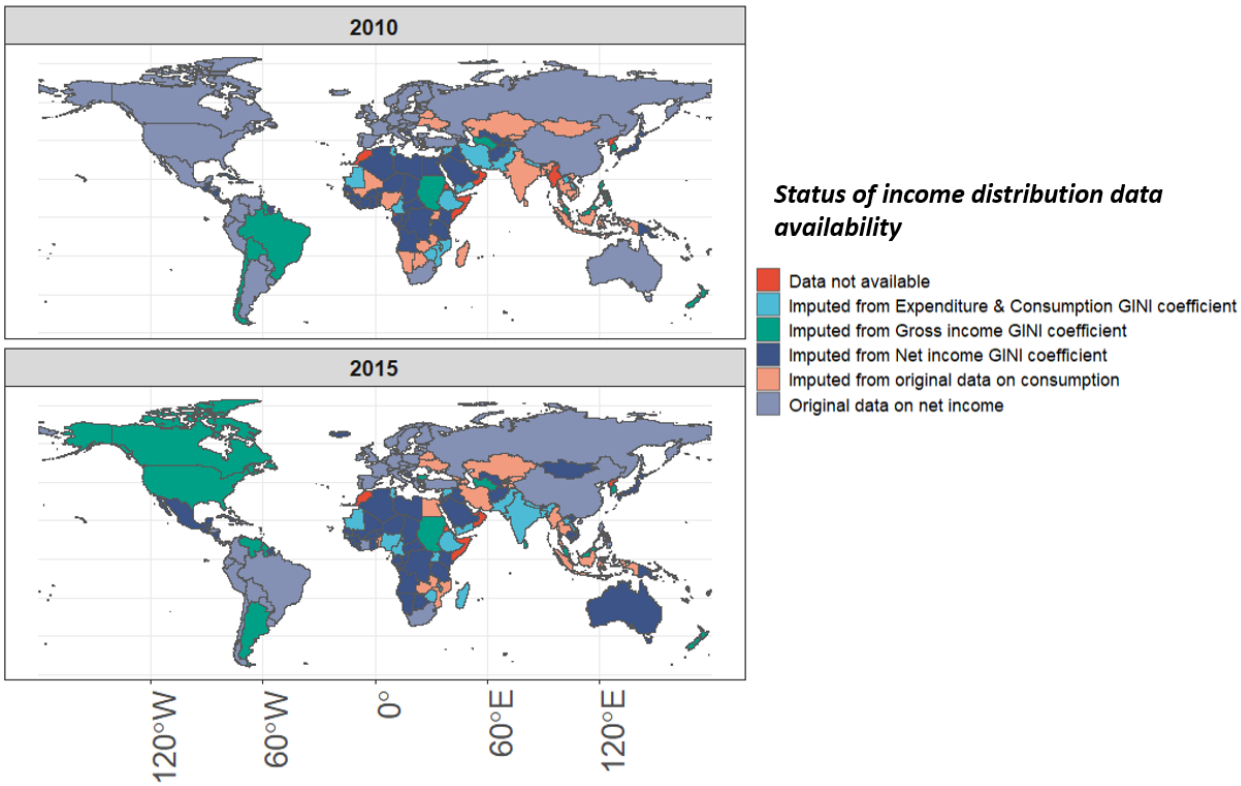

Figure 1:Data availability by country. Availability is shown here for two years 2010 and 2015.

iv.) We stopped our dataset in 2015 since that is the last historical year currently in the GCAM model. We can and should extend this dataset to other years in the near future as more data become available. This is now noted in the discussion section of the revised manuscript. We have made our code available, so this update should be straightforward.

v.) Our final dataset is a panel without gaps. For each year, we note where the data came from (original data vs imputed data from consumption vs imputed data from GINIs)

2. "The clear distinction of net, gross income, and consumption is a great contribution of this article. but not sure if I would call it "income concept" as consumption is not income. Rather a concept of inequality, income or consumption. But just my thought here."

Response: We agree with the reviewer regarding this point. We would note that the references to the "income concept" are available in the literature, especially the cited papers such as Deaton & Zaidi (2002). However, we have edited the abstract (Line 12-13) to introduce the "concept of inequality (or income concept)". Similarly, we have defined this as

the concept of inequality but mentioned that we will call it an income concept for convenience up front (on line 4-5 Page 2).

3. "Three notably sources of inequality data I see not references here, so a paragraph on those or why or why not they are included would be good to complete the picture: The SWIID database, the WIID database (which might have been used, but the name is not given), and the WID by Chancel et al. (All three databases can be easily found online)."

Response: We have now addressed this in section 2.1 (Literature review) on lines 1-8 Page 5. We are primarily interested in decile-level income distributions derived from household surveys. Given our criteria for data selection, we limited our data collection to the datasets mentioned in the manuscript (LIS, PovCal). We did not use the Standardized World Income Inequality Database (Solt, 2020) since it includes only the GINI coefficient and not a full distribution by income groups (such as deciles). Moreover, the concept of inequality in this dataset is disposable income or market income as opposed to net-income. Similarly, we did not use the World Inequality Database (Chancel et al 2021) since this dataset is not based on household survey data (This database uses a distributed national account methodology). Finally, we confirm that we have used the WIID database. We refer to it as the UNU-WIDER database. This is now clarified in the manuscript.

4. "Equivalence scale: this is an important feature of income and inequality measurements. Some more detail would be good. You argue on page 6 to use "same per-capita equivalence scale". First, it is not clear what this means, but you seem to divide HH income by all members, counting even children by 1. Are many surveys not use say the OECD modified or square root or similar equivalence scales? This is an important issue in my opinion to at least discuss."

Response: Thank you for this comment. We have clarified in the manuscript (Page 8, Line 9-11), that we selected observations that were all represented at the same per capita equivalence scale. The UNU WIDER dataset presents data across sources at the same scale and in fact allows users to select from amongst 3-4 options. We selected the per capita scale since it was the most appropriate for GCAM. We have not tested the effect of changing the equivalence scale (e.g., OECD vs per capita) on the income distribution, however. This can be explored separately in future studies.

5. "Sections 2.x should be slightly improved in terms of presentation. In 2.3 the regression estimation would be good. The sentence "We calculate values for 9 deciles d1-d8 and d10 and the re-calculate d9 as the residual." needs to be explained. Why d9 as residual? Equation 1 is a projection but should be a regression estimation in my opinion. And please harmonize notation with the PCA section. Would suggest beta for "Coeff". etc.."

Response: Thank you for this comment. We have made several corrections to this section. Namely-

- **i.)** **The notation of equation (1) has been changed and the parameters are presented similar to the PCA equations later in the manuscript (Page 12).**
- **ii.)** **We have clarified that this equation is a regression equation (Line 6-7 Page 11). This has also been clarified by moving Table 4 up so that it is clear that we used 10 separate equations for each decile.**
- **iii.)** **Finally, we have clarified that d9 is calculated as the residual since it had the lowest R squared (0.43) amongst the 10 regression equations for the deciles (Line 1-2, Page 11).**

6. *"You refer to another paper of yours (Narayan et al. (2023)) describing the method of PCA in detail. But that paper seems not to apply to historical but future scenario data, correct?"*

Response: Yes. This PCA model was indeed used to generate future projections. However, this was fit to the latest available data on net-income distributions (only original data with no imputation) and was found to provide a better fit compared to other imputation methods such as the lognormal functional form (See Figure 5 in the manuscript also attached below). Hence this method was used to impute historical values where only the GINI is available. This is now made clear in the Introduction (Line 11-13 Page 3) and in section 2.4 (Line 22-25 Page 14).

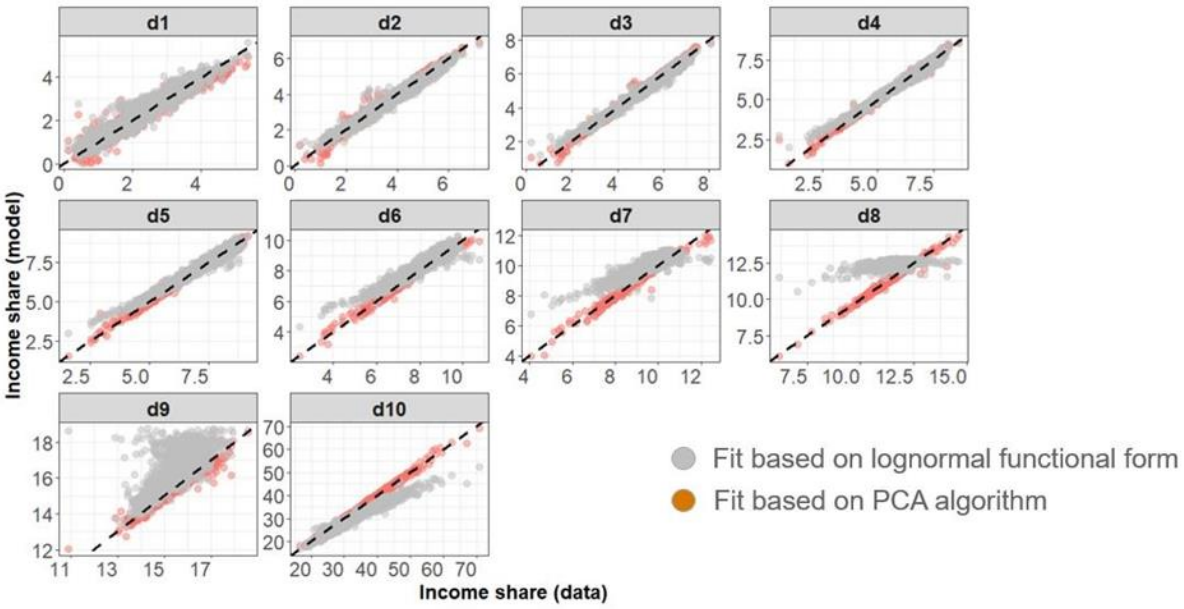

Figure 2: Comparison of fit of lognormal functional form (grey dots) with PCA based fit (orange dots) with data for each decile (facet). Lines represent 1 to 1 fit between x and y axis. Income shares are expressed as a percent of total income.

7. *"One bis issue with inequality based on survey data is that surveys in many countries are run only every x years, so the time gap is one of the big issues, yet this is not mentioned specifically in the paper, might be good to add. And related to this, in the cases where an aggregate inequality index was available, how had that been computed, as there must be some underlying distribution and hence deciles available."*

Response: Thank you for this very thoughtful comment. We have not addressed the issue of variability of surveys between years. However, the LIS and the PovCal have consistent methods in each year which is a big advantage of this data. We used the WDI GINI to fill in holes. But we have noted that the WDI does not provide any details on underlying surveys. As a first step, we used the All the GINIs dataset to at least understand when a GINI is based on consumption, gross income or net income (These points are documented in section 2.4 now, on lines 13-16 Page 17). But as more data become available, this should be addressed more directly. We have added these points to the discussion.