Summary

This paper introduces a global reference land cover dataset at 10 m resolution based on Sentinel-2 imagery from 2015, containing over 16.5 million data records across 12 land cover classes. The dataset was created through expert visual interpretation of high-resolution imagery (e.g., Google Maps, Bing, ESRI World) along with additional sources from the Geo-Wiki platform such as NDVI time series, Sentinel-2 time series, and geo-tagged photos. The dataset is publicly available via Zenodo and supports applications in land cover analysis, ecosystem modeling, biodiversity, and cropland studies.

Major comments

The paper is well written and concise, the methodology is rigorous and sound, the study will contribute to the land cover and land use change community. I see great potential for publication in ESSD. However, there are several shortcomings and clarifications that I strongly suggest the authors address prior to publication. For example, it's unclear from the manuscript how misclassification was determined and how quality of reference data was assessed (see specific comments below).

Minor comments

- Table 1 – the term subpixel is not defined. Is a pixel 100 m and subpixel is 10 m? Please clearly define the term in the table. I see the definition in the text on line 73. Note that "subpixels" is spelled inconsistently throughout the text – sometimes it's spelled as sub-pixels and other times as subpixel.
- Line 22 – can you elaborate on how this can be used for biodiversity? It's not obvious to the reader.
- Line 82 – I think the authors mean Google Earth Pro and not Google Earth Engine as Streetview and historical imagery are available on Google Earth Pro.
- Section 2.3 – can you make it explicit that the visual interpretation was done for the year 2015? Could you also elaborate on how you used the land cover maps – I am assuming they were used as ancillary evidence and were not sufficient on their own for labelers to decide? Otherwise, the labels might be reproducing errors in existing land cover datasets. Out of curiosity, was each sample interpreted once?
- Line 93 – how many interpreters were trained? Was this done through a crowdsourcing campaign or were the labelers employees at IIASA/university etc?
- Line 95 – it's unclear if the group of experts is separate from the interpreters trained. Is that a subset of everyone trained? Did experts serve a different function such as reviewing interpreted labels or they were interpreters themselves.

- Line 109 – how was it determined that they were misclassifications? Did a second interpreter check (agree/disagree)?
- Figure 2 – indicates wetlands as herbaceous while Table 1 defines wetlands as either herbaceous or woody. Can you clarify the discrepancy?
- Section 3.2 – could the data be used for a fractional cover classification? Maybe you could list that as a use case as well. Usage in bullet point #2 goes against the good practice for accuracy assessment/validation that has now been widely accepted by the remote sensing community. Maybe instead you could suggest the data be used for a statistical cross validation during the model refinement stages of analysis. https://www.sciencedirect.com/science/article/abs/pii/S0034425714000704
- Lines 146-151 – According to Table 1 "subpixels were classified as trees when trees fall in the center of a subpixel (10 m x 10 m)" and in this portion of the manuscript you are saying "In such cases, tree cover was not the dominant class within individual pixels, yet we still needed to label some of them as "trees" to match the overall percentage." Two things: 1) those two statements are in contradiction, 2) it's unclear what "to match the overall percentage" means, 3) up to this point the impression was that the labeling was done at 10 m resolution and now it appears it was done at 100 m and matched down to 10 m somehow. Can you please clarify? How was 65% cover estimated if not by determining how many of the 100 10 m pixels had tree at the center of the subpixel? This statement at the end seems to contradict the definition in the table.
- Line 155 – still unclear to me how misclassification was determined. See comment above – was it misclassification relative to another interpreter or expert? This is important as you claim that this is a high-quality dataset but it's not exactly clear what metrics were used to determine quality.

Zenodo comments

- The difference between validation_id and sampleid is unclear from the description. Seems like they are the same thing.
- Based on Table 1 I thought unique_id would be a value between 1 and 13, however, the values are totally different (e.g., 3027, 3024) and not described anywhere. I think it will be useful to keep these consistent to help users plug this dataset directly into their analyses (which usually require numerical values for classes).