Review Hyde 3.2

Very important resource, prominent in, for example, other ESSD papers (carbon, fertilisers). This particular version, 3.2, already in use - does the adjective 'New' really apply? Whether 'new' or recent, a large community will use this description and access this data! Good match for ESSD.

Better presentation would make it more useful; see several suggestions below. I also worry about the supplement because, as I understand, ESSD does accept supplements because their parent publisher, Copernicus, does not archive supplements? My suggestions below include recommendations about dividing material in the present supplement between the manuscript and the data archive.

Not my intent to re-write this entire manuscript, nor to request the authors to re-write, but at least the introductory paragraph of the abstract should provide better, more concise description:

"This paper presents an update and expansion of the History Database of the Global Environment (HYDE, v 3.2.000). HYDE is an internally consistent combination of historical population estimates with time-dependent maps for land use, covering the time period 10 000 BCE to 2015 CE. Population is represented by maps of total, urban, rural population and population density as well as built-up area. Land use categories include cropland, with a new distinction into irrigated and rain fed crops (other than rice) and irrigated and rain fed rice. Also grazing lands are provided, divided into more intensively used pasture and less intensively used rangeland. This version applies enhanced land area allocation algorithms with improved weighting functions, adds the new land use categories described above, and takes advantage of recent information sources including newly-available census data." *(Add other improvements to this list? Tell the reader / user what is different or better about this version!)*

Although this reader found the narrative parts of this manuscript - particularly the historical accounts of the parallel developments of population and agriculture - quite interesting, I also wondered whether those accounts actually buttressed the data itself or instead provided an interesting but not-really essential distraction from the data themselves. I raise this question in several locations, below.

Throughout: replace the word 'weighing' with 'weighting'. Decide 'rain fed' or 'rain-fed' or 'rainfed' (this reader prefers the latter).

The manuscript contains frequent small grammar errors, probably a function of writing in English. I assume final proofreading will identify and correct those errors so I did not note them here.

Final assurance to the authors: by the time you reach the end of my comments below, you will sense a disappointment tending toward irritation in response to this manuscript. Please hear this assurance: I believe this data set has great present and future impact for a wide variety of uses! But, in proportion to its utility and impact, I think it deserves and needs a much better presentation.

Page 1, line 20: Define 'cap'. Also this number, 56 cap per km, represents population density per (potentially) live-able land area (e.g. excluding oceans, Antarctica, etc). Does this audience need that distinction expressed clearly at this point?

Page 1, lines 23-26: cropland 2.3% in 1700 = 294 Mha while rainfed amounts to 2.2% and also 294 Mha. Rounding error? More uncertainty in the total cropland estimates than in the distinction rainfed vs irrigated? Irrigated 5 Mha plus rainfed 294 Mha exceeds total cropland of 294 Mha in 1700? Here we first encounter one of the weaknesses of this presentation, the absence of quantitative uncertainty estimates.

Page 2, lines 3,4: If the authors intend to include these population and land use paragraphs in the abstract, then perhaps this data access line should go earlier, at the end of the first paragraph.

Page 2, line 7: This opening sentence refers to human impact over a millennia but the discussion in the paragraph that follows covers 1000 millennia? One could more accurately start with the statement 'Humans have emerged as the most important driving force of landscape transformation.'

Page 2, line 13: The terms 'Industrial Revolution' and 'Great Acceleration' refer to different time periods and different human impacts. I recommend to leave these labels out entirely as they do not assist the user to understand the data.

Page 2, line 19 to Page 3, line 6: This justification section for accurate land use data wanders through ESMs and introduces many acronyms not used later in this paper. I recommend to delete both of these paragraphs. The authors could very easily, and probably should, add an 'Impacts' section after the discussion that could concisely mention the relevance of HYDE 3.2 to many of these communities (carbon budget, ESM models, fire records, biodiversity inventories, etc.).

Page 3, lines 13 to 19: This paragraph repeats verbatim what the authors wrote in the abstract (which I suggested to modify). Again here we should get a very clear sense, perhaps even with a bulleted list, of the improvements of 3.2 over 3.1. Tell the readers what you have accomplished with this new version!

Page 3, lines 22, 23: Repeated information. We don't need these two introductory sentences.

Page 3, line 25: Suppl. Material Table S1.doc should go as an appendix to this manuscript. It contains mostly repetition and white space. I believe the authors could reduce its 'use of space' by at least 80%. Or combine it with other data as a spreadsheet or as a tab in spreadsheet. Other ESSD products, e.g carbon budget, make very effective use of spreadsheets to present country data.

Page 4, line 5: Supplementary material Box S1 should definitely appear as an appendix to this document.

Page 4, line 9: define acronym FAO on first usage. (The authors need to follow this practice throughout the manuscript.

Page 4, line 14: which curvilinear trajectories were applied to which countries and why? Based on data availability or non-availability. Based on development pattern? Geographic location? What and why?

Page 4, line 20: "specific rules for the deep past." What specific rules? Specific per country? Specific for each development phase? Some hint of this appears later, on page 16, but a reader or user never gets a clear statement of what assumptions applied to which countries or time periods. Apparently we need to consult Klein Goldewijk et al. (2016) - see page 17, line 12 - but no such reference exists in the reference list?

Page 4, line 28: "Supplementary material Table S1" - as above, reduce the space and make this an appendix or a tab in a data spreadsheet.

Page 4, line 30: New ESA land cover maps from 2016 - this is part of what is new in 3.2? Add it to the list (above)!

Page 5, line 7: First reference to Table 1. Agree that table 1 contains necessary information but it should be visibly much more informative and easy to read. Use dividing lines, colours, bars - many examples exist about how to present tabular information more effectively.

Page 5, line 11: Note reference to Waldner et al., 2016. Is this another new feature of HYDE 3.2? If so, list it?

Page 6, line 2: Supplement Table S4, 'data' sheet, - this should definitely go as part of the data set itself.

Page 6, line 10: Note reference to Siebert et al. (2015) - this is again part of what is improved in 3.2? Add it to the list?

Page 6, line 16: Supplement Table S2 - definitely belongs in the data set.

Page 6, line 17: MIRCA? Either define the acronym or omit it and simply refer to Siebert.

Page 6, line 23: Interesting logic here. Do protected areas represent land now protected after prior agricultural use, or protection of lands now that in the past, because of inaccessibility, never had agricultural use. More of the latter, I suspect? In which case this sentence is not correct?

Page 6, Section 3.2, Allocation of Cropland: Figure S1 should go here, very useful in this location.

Page 7, line 9: A much more recent version of the CRU - that acronym needs a definition - exists, more recent than New et al. 1997, although it might not change the 1960 to 1990 baseline period (but perhaps that itself needs revision both more recent and longer in the past?). See for example Osborne and Jones in ESSD 2014?

Page 8, line 2: Figure S1 should go in this manuscript (see comment above) and Table S3 should go with the data set.

Page 8, line 6: define IFPRI

Page 8, line 17: Table S4 should go with the data set.

Page 8, line 28: But, as already noted, within this qualitative geographic constraint, e.g. that all irrigated land must fall geographically within identified cropland boundaries, there are no quantitative constraints? As already noted, in introduction 5 Mha irrigated plus 294 Mha non-irrigated does not match to 294 Mha total cropland for 1700? Later, page 14, we find irrigated cropland at 6 Mha in 1700! I don't dispute the numbers, perhaps they result from rounding errors. But we need some quantitative estimates of the uncertainties of these numbers? As presented, all terms - population and agricultural - across all time periods seem to have the same and quite good numeric precision. That can't be true?

Page 9, line 1: Same comment about New et al. 1997 replaced by later (better?) data. Table S5 should go with the data set.

Page 9, Section 3.2.4, Grazing Land - Figure S2 should go here.

Page 9, line 15: define BIOME

Page 9, line 28: Table S6 should go with the data set.

Page 9, lines 32, 33: Do the authors have any basis for these numerical distinctions between open rangelands and managed grazing lands, e.g. the aridity index and/or the population density. I think they make a valid distinction here but I do not understand where the numerical basis comes from.

Page 10, Results: here we start the historical narrative which accompanies the data presentation. I agree that we need some explanations of what the data shows in trends of population and land use

but occasionally this section wanders into speculation? I agree that such a narration fits better here than in the Introduction. The authors should ensure they have only the minimum necessary information here. The editors may need to consider the appropriateness of that narrative information for the data journal.

Page 10, line 10: do we need this paragraph?

Page 10, line 19: agriculture very dependent on climate - what indicators does this data set have that support this statement? Do the authors mean 'dependent on local climate (environment) conditions' or to changes in large scale climate?

Page 10, line 28, Figure 1: Figure 1 continues very useful information but is very poorly presented. It contains mostly white space. The scales on the vertical axes change for region to region - the figure caption should point this out. Talk to a scientific graphics expert? This figure could be much more effective!

Page 10, line 30 to page 11, line 20: These paragraphs are not essential to describe this data set? I enjoyed reading them, but one could summarise in a short sentence that demographic and socioeconomic economic changes led to a global population of 253 million … as the starting sentence of the paragraph starting at page 11, line 22?

Page 11, line 25: Table S7 should appear as part of the data set.

Page 12, line 4, Table 2: Useful information but could have a much much more effective presentation. Change to landscape aspect ratio, add lines, colours, boxes. Need to be much more effective!

Page 12, line 23: Put Table S7 in the data set (already mentioned). Table 3 draws the same comments as Table 2 (above) - make it much more effective, put it in landscape aspect ratio, at lines and colours, etc.

Page 12, line 25: These authors mention differences to Houghton 1983 and some possible factors (tropics, pasture vs rangeland) but do not provide a summary. The reader does not get a sense of the basis for confidence in the present data?

Page 13, line 1: Table 3, see improvement suggestions above.

Page 13, line 3: 3000 CE?!?!

Page 13, top paragraph: abundant opinion here but not much quantitative explanation?

Page 13, starting line 14: Nine sentences of historical context here followed by one sentence of data? Do we need all the preface?

Page 13, line 25: Table 4, again, needs substantial visual and organisational improvements to become much more useful. I question the apparent precision of these numbers as presented?

Page 13, final paragraph: I like this explanation but do we need it in the manuscript to properly describe these data?

Page 14, line 1: Figure 2: Useful information but poorly presented. Needs much more detail in the figure legend. Needs visual improvements! For example, in right panel (1500 to 2010CE) we have, apparently, 8 data sources other than HYDE 3.2 to compare. But at best we can only see 3?

Page 14, line 3: here global irrigated area = 6.0 (! such precision?) Mha where earlier we saw 5 Mha? Uncertainties of both of these numbers?

Page 14, line 4: “figure xx” ????

Page 14, line 8: figure 2 does not show country data!

Page 14, starting line 10, Rice: Two (interesting) paragraphs of historical context before we get to the actual data! Do we need those paragraphs??

Page 14, line 26: here uncertainty emerges explicitly, but apparently related to weakness in assumptions about historical rice cultivation in China? How does this specific uncertainty contribute to overall uncertainty (see below)?

Page 14, line 30 continuing to page 15, line 9: Two additional (still interesting) paragraphs of historical context before the next paragraph of data - does the reader need these paragraphs to better understand the data? Perhaps the authors believe so, but they have not made a persuasive case? I suspect most users do not need these accounts?

Page 15, line 10: Figure 3, not quite useful and visually not interesting. Very hard to distinguish the various data sources. Need colour and better symbols. Why not colours as in Figure 4?

Page 15, line 14: Very sloppy use of shorthand, e.g. 0.20 Mha w (0.20 t). ‘w’ means ‘wet’ and ’t’ means ‘Total’?

Page 15, line 15: In fact, a reader has a very hard time to distinguish any of these data in figure 3?

Page 15, line 22: Table 5, as for all other tables, needs a much different and much more effective presentation. Same suggestions apply: landscape aspect, colours, lines, boxes, etc. Again, what should we make of the precision in these numbers as presented?

Page 15, lines 28 and following, Summary: A reader might agree with these statements about improved yield and global nutrition but the data in the paper, and in particular figure three and tables 6 and 7, say nothing whatsoever about yield or productivity. These data only present land coverage areas for agricultural use and global population. I don’t believe - based on these data - that the authors can make these statements!

Page 15, line 31: figure 3 does not belong here, the authors already introduced it at the top of this page.

Page 16, top paragraph. The authors seem mystified as to why differences so large occur between this data and Kaplan. We get no additional information. Could the Kaplan algorithm of deforestation with population have serious assumption errors, or serious errors over time? We need more information here?

Page 16, line 7: Table 6 necessary but needs better presentation. Could this work as an X-Y plot instead?

Page 16, line 11: Table 7 necessary and useful but needs a better presentation

NOTE: Other than in the figure legend itself, I find no mention of Figure 4 anywhere in the text?

Page 16, Uncertainties

This reader finds this uncertainty section unhelpful and unsatisfactory.

I would like a clear expression:

Uncertainties in the population estimates, how those vary with time, and how other data sources, tools or interpretations might remedy those uncertainties; then

Uncertainties in agriculture land use categorisations, how those vary time, whether cropland or grazing land distinctions induce more or less uncertainty, likewise whether irrigated vs non-irrigated induces uncertainty, and what tools, data sources or other approaches might mitigate these uncertainties; summarised with

An overall assessment of uncertainty, perhaps as a percentage probability or as a plus-minus standard deviation.

The present information ranges from very qualitative in terms of population ("not … unfair", "acceptable", "feasible") to apparently quite quantitative (e.g. the fuzzy numerical Kappa analysis) but applied only to three land use snapshots including a recent ESA satellite product but not a HYDE 3.2 product (yes, we did earlier get assurance that HYDE 3.2 converges to ESA 2010 in the present). Figure 5, with A and B uncertainty ranges appear to represent only human estimates applied with simple uniformity across time, e.g. factor of 1, factor of 2, etc. Other than the comparison of 3.1 with 3.2, a reader really gains no useful information from Figure 5? Perhaps the uncertainty ranges reduce substantially in the recent data-rich years, but we can't see that in Figure 5. Or, instead, the authors have simply applied a plus minus 5% to present day data, increased that arbitrarily going backwards (earlier) and then doubled it for range B.

I appreciate that the authors work very much at the edge of what we can say about deep historical trends of both population and land use. I also respect, and would tend to accept for my own work, their expert judgement. But if so, then apply these 5% and 10% uncertainties consistently throughout the paper and data instead of giving us two significant figures in tables of land use areas?

Finally, following a suggestion earlier, if the authors move a revised uncertainty section before the summary, then a definitive summary could list the many possible uses and users of this data (e.g. carbon or nitrogen budgets, ESM modelling, fire records, biodiversity indices, etc.).