Review ESSD-2020-159

Sorry for delays. I made a long careful effort to read Descals et al. and to re-read Xu et al. https://doi.org/10.5194/essd-12-847-2020.

I appreciate seeing these products emerge via ESSD. Our land use / biodiversity / global ecology community should welcome solid descriptions of, and open access to, remote sensing products. Both papers should help us understand what works, how to use it, but also what doesn't work. For those confronting complicated land use, social and economic issues around expansion of oil palm production and use, both these data products should prove valuable. When I first read regional (Malaysia & Indonesia) study I suspected and hoped someone would extend those tools to a global reach. Because we could soon have two products of similar intent but using substantially different approaches, I anticipate positive research effort to sort these differences. I also understand how Google Earth Engine (GEE) has stimulated and facilitated this study but I remain very cautious as our remote sensing researchers develop greater and greater dependence on Google.

About the specific product here, I have many quibbles, questions and suggestions. Please understand that with most of these comments I try to make a good combination of remote sensing and machine learning accessible to larger audience of ecologists and concerned citizens, to enhance what these authors call (line 434) "discussion about environmental impacts of oil palm, including on biodiversity". Overall, we definitely need it and - due to high attention to quality - ESSD seems like a good place to publish it. A host of changes and improvements will make it much more useful.

I request: a) substantial change/improvement in identification and specification of sources and codes; b) extended explicit explanation of uncertainties; c) justification and explanation of statistical approach; and d) better discussion of how various definitions used here ('closed-canopy', 'industrial', 'smallholder') impact both this work and external comparisons. I follow with a list of specific comments, most of which echo overarching concerns.

A) *Source information and documentation*. Following journal expectations and good practice for supporting future users, please provide exact detailed information about all source products, whether from GEE, from your own work, or from external publications. Provide sufficient information so that a researcher not using GEE can reproduce your exact steps: product name, DOI, version number, other essential metadata (scan mode, resolution, date), date of last access, source of last access, URL, other references, etc. Most users will not attempt to reproduce your work but you must provide sufficient detail so that they could if they choose. Assure us that all sources remain free and open access to us as ordinary users. The issue of GEE looms very large here. I understand convenience and advantage of GEE. I use GEE, Google Earth, Google Scholar, Gmail, etc. for most of my daily work. Undeniable strengths and capabilities of Google products do not include one factor essential for data exchange and data publication: reproducibility. Google makes frequent substantial improvements in products and services, almost always in a manner hidden to users and absent any provenance. GEE may pride itself in providing latest version of a given LandSat or Sentinel data product, but woe to a user who tries to track history of that product. Likewise for imagery in Google Earth or search outcomes in Google Scholar. From a data reproducibility view, nothing seems reproducible, certainly not on annual time scales and occasionally not even on weekly timescales. Of course we must use holdings and computational resources of GEE! We must also, however, perhaps following the political adage about 'trust but verify', provide complete accurate documentation of all sources used within GEE besides those assembled outside by our own efforts; future users encountering changed services from GEE must have the ability to obtain same sources

reliably elsewhere. Data in most cases derives from original (e.g. ESA, CDS, whatever) sources, as authors hint at line 417. I recommend authors provide a table, e.g. Table 2 in https://doi.org/10.5194/essd-12-1217-2020, listing all sources with all necessary metadata (DOI, version, reference, date of access, etc.). For this paper such a table would include at least Sentinel-1 SAR, Sentinel-2 vis, DigiGlobe imagery, IUCN base layer, the exact CNN codes (DeepLabv3+, MobileNetV2). Authors no doubt hold all this information among themselves, and cite (most of it) appropriately. Please put it in a single easy-to-use recipe table for users.

B) *Uncertainties*. From abstract throughout the entire manuscript, these authors portray area precision of four significant figures: 17.47 x 10^6 (line 29 and many following); and statistical precision of three significant figures: 85.6% (line 28 and many following). Whether a reader / user accepts or doubts these precisions (this reader doubts), authors provide zero basis for uncertainty assessment. Quoting from ESSD guidelines (https://www.earth-syst-sci-data.net/10/2275/2018/): "every ESSD data product must include uncertainty documentation." Nothing here. Authors have not documented whether or not they can defend any areal estimate to better than 10^6 hectare, or any kappa value to better than 10%. If authors find this comment offensive, defend it! You have - at minimum - uncertainties in source data, uncertainties generated by CNN models, uncertainties in definitions (e.g closed-canopy, discussed but never quantified), and some combination of these and other uncertainties in your final estimates. Not once does a user find a 95% CI, a standard deviation; nothing. Authors provide no confidence basis for comparisons with Xu 2020, with Gaveau in prep, etc. I start from an assumption that these authors have substantial skill with and command of uncertainties introduced by their approach and their tools, but they fail complete to justify any of it. Absence of quantitative uncertainties contaminates their extensive distinctions and discussions of industrial vs smallholder; how does a reader credit any of that if we lack confidence that they can even determine those areas to better than $\pm$ 10%?

C) *Statistical approach*. Presuming the author list includes expertise in machine learning, the authors do nothing to help a larger reader community understand their approach. A reader gets no explanation of accuracy and kappa coefficients. Authors assume that readers will automatically understand "user" accuracy as commission errors and "producer" accuracy as omission errors. If, as you hope, this work stimulates wider attention from global biodiversity community, make sure they/we know what you are talking about (and know that you know what you are talking about). Explain true positives, false positives, false negatives, etc. You present clever data manipulation (supplement figure 3): what exactly to you test with this manipulation? By how much (if at all) did that manipulation increase skill or reliability? Provide tangible examples of positive and negative outcomes, in oil palm terms of area, age, type of planting, stability of land use patterns, etc.

D) The *data approach* described here, with an undeniable focus on mature orderly large-scale industrial closed-canopy plantations, provides (or could, if we knew uncertainties) one valid way to monitor and inventory oil palm distributions. By the authors' own admission, their "… analysis generally does not detect young oil palm", "struggles to detect oil palm in non-homogeneous settings", and, as a consequence "is an underestimate" (lines 352 to 354). Fair warning, well taken. Despite caveats however, the authors would have us accept these data in preference to alternate data products (using similar space-borne radars to monitor clearings over time as in Xu et al.) or to a future as-yet un-described product from one of the co-authors relying on LandSat (e.g. non-radar) images? They should certainly describe merits and limitations of their work for benefit of users! Reader encounters statements like "Comparison with Xu's data requires a degree of caution because it only reflects the accuracies for closed-canopy oil palm plantations, while the multi-annual analysis in Xu et al., 2020 aimed to detect disturbances in the time series in order to classify young plantations" (line 302). Caution indeed, but by whom? Apparently, by us as readers, but never-mind that these authors have given us too little information to accept certainty or

uncertainty of their work. They have strayed from data description to (unsupported) advocacy? In fact, with their help, we might have a very useful piece of the puzzle here: a definitive snapshot of closed-canopy industrial-scale oil palm distributions and areas (definitive includes quantitative uncertainties, sorry to harp on that theme). From that newly-available starting point, then a) venture tentatively but provocatively into industrial scale vs smallholder scale issues, b) compare with FAO production numbers, noting matches and mis-matches, and c) validate or not with alternate approach of detecting land use discontinuities.

Present your best case for the closed-canopy product. Show how you have applied new tools to an urgent challenge. Build reader/user confidence in accessibility, reliability, traceability, reproducibility, etc. Then show how your new product changes the picture, e.g. suggests new directions. Authors need to change/improve presentation to help readers better understand their accomplishments.

Specific comments

Line 100: In addition to useful workflow in Figure 1, give us a table with details and documentation of all data sources and tools (comment A, above).

Line 105: Ditto comment A. Is WorldClim V1 Bioclim openly accessible? Validated? Available via GEE?

Line 107: Ditto for IUCN.

Line 110: Who compiled LandSat images in GEE? Cloud-free images for one year (2017) over the tropics? One suspects only a few images, perhaps not at optimal seasons, per location? Curiosity or actual relevance to this work? Authors give us no clue.

Line 113: “Supplementary to this, Table 1 shows the minimum …” Is this an oblique way of referring readers to STable 1? Confusing? STable 1 comes from these authors, IUCN, someone else’s work?

Line 115: Implies bioclimatic information used as qualifying criteria, e.g. seventeen of nineteen needed. But SFigure 1 shows full range (0 to 19) of qualifying bioclimatic variables, nothing about any cut-off of meets or does not meet. Reverse scale to most attribute maps: max number indicated by absence of color/shading?

Line 124 (Legend to Figure 2): where closed-canopy oil palm land use was detected by CNN?

Line 133 - tell us the wavelength (665 nm?) of Sentinel-2 band 4. Especially because you later talk about LandSat but wavelength of LandSat Band 4 does not equal wavelength of Sentinel-2 Band 4.

Line 140 - did you use NDVI or fixed Band 4 wavelengths. NDVI implies red plus near IR (usually 665 with 810 or 840) but here so far you have only talked about 665. If you used an NDVI product, did you recalculate it from sensor wavelengths or use a canned corrected product available from e.g. Sentinel-2?

Line 145: roads, mills. Roads we hear more about but in regular vs convoluted rather than presence/absence terms. Mills we never hear more about? Part of a black-box approach to CNN? Assume the algorithm finds them useful because somebody else reported same?

Line 147: high-res images from DigitalGlobe - accessible via GEE? Provided separately by authors? We need an exact data table of all source materials!

Line 156: 96 altered images. Nicely illustrated by SFigure 3, but so what? Reader never learns how or even if this manipulation had any impact on image discrimination, on commission or omission errors, on uncertainties, etc. Clever, but for what quantitative purpose?

Lines 158 to 179, introduction to industrial vs smallholder oil palm land management. If authors demonstrate that their product supports quantitative reliable discrimination of industrial vs smallholder, then the introduction here, addressing properly the host of economic, social, environmental and political complexities around small vs large, seems only mildly redundant with later discussion e.g. at lines 360 and following. If, however, by rigorous uncertainty analysis, the authors show that they can not distinguish what they call smallholder from what they call industrial to better than plus/minus 10% or even 20%, then this section seems entirely premature, better included in a later discussion of what they can and can't say about smallholdings?

Line 185: "CNN can automatically learn contextual information such as the road network in industrial plantations …" Crux of the issue. In a good data description, authors would attack such a statement, doing their best to show how they assembled sources and defined terms to maximize quality and reliability of CNN outcome. Instead, here, we get a sense that they anticipated strengths of CNN and 'let it run'. Is the specific CNN code included in GEE? DId they use GEE computational resources or their own computational resources to perform this analysis? How much uncertainty did they / we face beforehand? Later they discuss how few training points they used compared to other uses, but never in terms of reducing or accepting uncertainties? How did the CNN increase that uncertainty or - one hopes - produce an outcome whose value outweighs increased uncertainty. If they only want to advocate machine learning via CNN as a potentially-useful tool for analysis of oil palm distributions, that paper should go elsewhere? They have a chance here to convince users of a real step forward using some new tools, but they have not yet given us confidence in the results. Reader needs more information about DeepLabv3+, especially in a source table. Chen et al. 2018 - obscure conference proceedings - not a useful reference.

Line 195: Validation. Authors show 10k+ validation points of which something like 6% represent valid closed-canopy oil palm plantations and something like 1% represent - by their definition - smallholdings. This reader does not know whether authors deliberately tried to prove efficiency at detecting rare land use types or conducted a truly randomized trial that unfortunately reproduced the rare occurrence of closed-canopy plantations? We would like to trust their skill, but they provide too little information.

Lines 198, 203: Part of the confusion above comes from authors reference to "high-resolution images" and "high-resolution images in Google Earth". What hi-res images? Available via GEE or selected separately by authors? 5 metre resolution? Better? What dates or seasons? I can't get a GE image more recent than 2014 for my current location; because GE provides no provenance I estimate the resolution as perhaps 1 metre. What conventional or special images did the authors use or obtain? Users need this information in the much-mentioned data source table?

Section 3, Results: Authors provide useful and nicely-illustrated section describing outcomes of their analysis and user-benefits of their new data product. Unfortunately, absent source documentation, quantitative uncertainties and better explanation of statistics, we - especially if 'we' includes ecologists, economists, land use specialists, biodiversity campaigners - can't credit any of it.

Line 319: “confirm previous findings on the suitability of radar satellite data for mapping” But this is not only a radar study; it includes visible spectrometry as well. Or, have supposed benefits of including Band 4 of Sentinel-2 disappeared by now. Or proved disappointing, unsatisfactory from a statistical standpoint? How would a user know?

Line 328: only the model will receive annual updates? Presumably source data also updates? If you change both CNN model and source data you will lose the ability to identify reasons for changed performance, e.g. better scene discrimination leading to better area estimates? If you want to detect a time rate of change, you will need accuracy, precision, and uncertainties so much missing here?

Line 342: “Gaveau et al. forthcoming” not useful. Frequently referenced but not helpful: no journal, no status. I do not know how ESSD or other Copernicus journals handle future submissions; I might have expected to see ‘pers. comm.’.

Line 360 and following, the entire discussion on lack of definitive global definitions for smallholdings. Possibly useful but authors would need to convince this reader before this point that their product had sufficient accuracy and low enough uncertainty to quantify closed-canopy first and then quantitively distinguish industrial from smallholdings. Absent that proof, one doubts that a this interesting discussion has any place in a data description.

Line 400, code availability. Much anticipated from prior references but the GitHub link takes one only to a java script code absent any documentation or explanation. As for all GitHub links the landing page starts with a version/update log, recent in this case but with no history.

Line 402, data availability. Easy data availability via Zenodo. The grid folder contains grid with and grid without overlays in several useful formats; the .shp files, although labeled as ESRI-format, open easily in e.g. QGIS (open source GIS software). .tif files in oil_palm_map files open in most image software.

Line 416: Example of deficient source information which should instead become part of detailed information in a source data attribution table. Confusing reference (in line 417) to access via GEE when GitHub link contains only a .js file.

Line 419, Conclusions. Premature at this point, will need substantial revision.

STable 4: Comparison to Xu et al., ESSD, https://doi.org/10.5194/essd-12-847-2020, but the header label for column 2 reads Chen, 2020. No Chen 2020 in reference list?