REVIEWER #1

This paper presents an up-to-date map of oil palm plantations by typology (industrial vs. smallholder plantations) at the global scale and with unprecedented detail (10-meter resolution) for the year 2019, and showed the suitability of deep learning in remote sensing for complex classification scenarios in which contextual information may be useful. It is supported by the knowledge available in the literature. The results showed that the method proposed is robust and repeatable. The paper is clearly described and the comparison with previous works showed better results.

L.38- as new imagery is published, it can be used to reliably to monitor the — as new

images become available, it can be used to monitor the

L.123- with green signifiy — with green signify

L.126- spectral images — optical images

L.126- satellites, both — satellites, respectively, both

L.187- This study emloyed the — This study employed the

L.384 -and FOA harvested — and FAO harvested

Dear Reviewer #1,

Thank you for your review. We appreciate your supporting words on the importance

and quality of the global oil palm plantations dataset.

We updated the manuscript based on your comments and recommendations.

L39 - ...as new images become available, they can be used to monitor the...

L145 - ...with green signify....

L147 - ...optical images....

L148 - ...satellites, respectively, both...

L217 - This study employed the...

L456 - ...and FAO harvested....

===

REVIEWER #2

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.

Dear Reviewer #2 - we truly appreciate your effort, time and willingness to comment on our work.

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.

Response: Thanks for your positive assessment on the interest and value of our study. We found that ESSD is indeed a good place to publish geospatial data; the preprint version has given publicity to the manuscript and we received some minor suggestions that go in line with some comments that the reviewer made. We have now addressed all your insightful concerns in the revised version. Please see below point by point.

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 ('closedcanopy', '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.

Response: We have now provided a table explaining all sources used in the study (Table 1 (line 120)). The table shows the product name and version (if any), the reference paper, and a short description about the way in which the data were used in the study.

We found it more appropriate to include the URL and details about how the data were accessed in the section *Data availability* (accessed via the official product portal or via GEE).

We have added some missing metadata in the main text:

- Revisit time Sentinel-1 (line 151)
- Revisit time Sentinel-2 (line 163)
- Instrument mode in Sentinel-1 (line 152):
- Central wavelength of Band 4 (line 155)
- Spatial resolution DigitalGlobal images (line 171)
- Availability of DigitalGlobe images (line 172)

We ensured that all relevant sources remain free and open access to users for reproducibility purposes. Previous oil palm datasets in Meijaard et al., 2018 and Descals et al., 2019 have been made publicly available and we provided the link to the repository of the convolutional neural network code.

We have explained now in lines 110-113 that Sentinel-1 and Sentinel-2 are the main data necessary to reproduce the classification of the global oil palm map: "*The Sentinel-1 and Sentinel-2 images taken in 2019 are the only data necessary to reproduce the results of the global oil palm map. The rest is auxiliary data used for the identification of the oil palm distribution, the visual interpretation of oil palm plantation, and the comparison with other oil palm maps.*"

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.

Response: As pointed out by the reviewer, we used GEE for convenience. Downloading and processing all the Sentinel-1 and Sentinel-2 images taken in 2019 for the area of study would have been a tiresome task that would require a high storage demand.

We primarily used GEE for creating the Sentinel-1 and Sentinel-2 composites. We also used GEE for the labelling of the training data and the identification of the potential distribution of oil palm, but these two steps can be easily implemented in other platforms. The bulk of the processing (training of the convolutional neural network and classification of images into the final oil palm map) was done with a local computer, not with GEE. We have included the Supplementary figure 1, which depicts the platforms and programming languages used in each step of the workflow and provides an insight about the actual dependence of this study on GEE.

We have also explained the programming environments used in each step of the workflow in the main text: (lines 103-110) "*The processing steps depicted in figure 1 were implemented in different computing environments (Supplementary figure 1) depending on the convenience for the processing. The compositing of Sentinel-1 and Sentinel-2 images was done in Google Earth Engine (GEE) (Gorelick et al., 2017) since a cloud-processing platform was suited for this task considering the high amount of satellite data required in the compositing. The visual interpretation of training data and validation points was also done in GEE. The training of the CNN and the classification of images, however, was performed with a local computer using Matlab 2019a since the implementation of the CNN model was less feasible in GEE. The CNN model can be also trained and used for prediction of images with Python (Code accessible through section 5. Code availability).*"

We understand the concern about reproducibility; GEE only mirrors the last version of a data product and users cannot access previous versions of a product in the GEE data catalog. For instance, the whole set of MODIS products version 5 were replaced by version 6 in the last two years. Despite this, GEE is very transparent about the version of the product, and users can check the version of the product from the code editor or from the website of the GEE data catalog: (https://developers.google.com/earth-engine/datasets/catalog).

Please note that GEE simply mirrors the data that ESA or other providers have delivered. We have added a paragraph explaining this in lines 497-502: "*When using GEE, the Sentinel-1 and 2 data are hosted and accessed in the Earth Engine data catalog [...]. Despite the data are hosted by GEE, these satellite data are the same as accessed via the official portal (Copernicus Open Access Hub: https://scihub.copernicus.eu/); Data ingested and hosted in GEE are always maintained in their original projection, resolution, and bit depth (Gorelick et al., 2017).*"

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 the 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.

Response: We thank the reviewer for these suggestions. Summarizing our previous comments, we have included table 1, which summarizes the data used in the study, and Supplementary figure 1, which depicts the programming environments used in the workflow. We also included in the main text a description of the programming environments (line 103-110) and a clarification about the data accessibility in GEE (lines 497-502).

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-scidata.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 ± 10%?

Response: Thank you for this useful comment. We have now provided an uncertainty assessment for both accuracy metrics and area estimates following the good practices for validating remotely-sensed data described in Olofsson et al., 2014. The revised manuscript now reports the accuracy metrics along with confidence intervals (confidence level of 95%). Figures and tables also provide the accuracy metrics with confidence intervals for our results, for the datasets presented in Xu et al.,2019, and for our previous study (Descals et al. 2019). With the uncertainty analysis, we could defend the claims regarding these comparisons.

The uncertainty analysis in the area estimates has been also useful to support the discussion regarding the comparison with Gaveau's dataset in Indonesia and FAO inventories in Western Africa, and extend the discussion about the uncertainty generated by the definitions of classes (closed-canopy vs. young and sparse plantations). The manuscript now presents both area mapped and area estimates, as introduced in lines 258-259: "*Here, we used the term area mapped for the total area classified as a given class, and the term area estimate for the estimation of the actual area following the practices in Olofsson et al., 2014.*". Our area estimates in Indonesia and Western Africa (now quantified with a 95% confidence level) are lower than the areas reported in Gaveau's dataset and FAO, respectively, inventories due to the definition of oil palm plantation.

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, 42-57.

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.

Response: Thank you for the suggestion. We have now explained in a comprehensible way the accuracy statistics used in the study (lines 244-254). These lines include an explanation about what the omission and commission rates signify from the perspective of a potential user of the oil palm map. Throughout these lines, we have repeatedly mentioned Olofsson et al., 2014, as it is a relevant reference in which the reader can deepen understanding of the validation practices in land cover mapping.

In addition, we have now included a new section in which we explain how we compared our results and existing oil palm datasets (Section 2.7 Comparison with other oil palm datasets).

Please note that we removed the kappa coefficient because this accuracy metric is highly correlated to the overall accuracy. We preferred to be more concise and present only the overall accuracy, and user's and producer's accuracy.

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.

Response: Thank you for pointing this out. We took for granted the improvement in model accuracy with data augmentation techniques. We have now justified the use of data augmentation (aka data manipulation) and developed the paragraph in order to justify the use of this technique: (line 181-184) "*Data augmentation aims to generate a more diverse training dataset with certain affine transformations applied to the original training data. Data augmentation techniques have been used in remote sensing studies (Yu et al., 2017), in which affine transformations such as flips, translations, and rotations have improved the accuracy results of deep learning models.*".

We have also cited a review paper that explains in more detail the data augmentation in deep learning, and cited a study that shows analytically the improvement in classification accuracy after such image transformations. We have also analysed the improvement in overall accuracy when adding these 96 augmented images for the validation subset in Sumatra. The overall accuracy was 94.02 ± 0.89% with data augmentation and 91.73 ± 1.03% without data augmentation (95% confidence interval). We did not include these results or deepen into details of this improvements in the manuscript for two reasons: (1) the improvement after data augmentation has been extensively referenced in a review paper (Shorten & Khoshgoftaar, 2019) and (2) such analysis would fit a deep learning journal instead, but here we considered that these results may distract from the main accuracy assessment and the comparison with other datasets.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60.

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 nonhomogeneous 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!

Response: The manuscript now emphasizes some of the advantages of using our oil palm map instead of other existing datasets. These are the merits mentioned in the Discussion:

- High spatial detail. Except our previous study (Descals et al., 2019), which only covers Riau province, other studies present a spatial resolution ≥30m.
- It is the first global product that depicts and differentiates between industrial and smallholder plantations.

- The uncertainty analysis shows that the accuracy for closed-canopy oil palm is higher than previous datasets (Xu et al., 2020 and Descals et al.,2019). Compared to Xu's map, we improved in terms of commission rate. Compared to our previous map (Descals et al., 2019), we improved in terms of omission rate.
- Gaveau's dataset for Indonesia delimited manually the plantations while our method classifies satellite images, which enables the monitoring of the crop in the future. Gaveau's dataset also mapped land likely cleared for oil palm but not yet planted, or failed oil palm, which provide different metrics than our current study on high yielding, mature smallholder and industrial scale plantations. This metric is important in discussions about meeting future vegetable oil demand given different land needs and yields of different oil crops.

As also commented by the reviewer, the manuscript warns the user about the limitations of the product:

- young and open canopy plantations, plantations in heterogeneous settings, and semi-wild oil palm remain undetected.
- The above implies that our area estimate reflects only closed-canopy oil palm and the actual area is higher if we consider all typologies of oil palm plantations.

We have also rewritten the abstract in order to incorporate these points in a concise and logical way.

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.

Response: The meaning of this sentence was not clear and we have rephrased it: (line 375) "*The comparison with Xu's data, however, only reflects the accuracies for closed-canopy oil palm plantations; the multi-annual analysis in Xu et al., 2020 also included the detection of disturbances in the time series in order to classify young plantations.*"

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.

Response: Following the reviewer's suggestions we have now updated our Abstract: the text points to the comparisons with existing datasets and FAO. We also included some more relevant results (areas and ratios), and it is now more compact, less repetitive and the logic improved, please see it below:

"*Our analysis also confirms significant regional variation in the ratio of industrial versus smallholder growers, but also that, from a typical land development perspective, large areas of legally defined* *smallholder oil palm resemble industrial-scale plantings. Since our study identified only closed-canopy oil palm stands, our area estimate was lower than the harvested area reported by FAO, particularly in Western Africa, due to the omission of young and sparse oil palm stands, oil palm in non-homogeneous settings, and semi-wild oil palm plantations. An accurate global map of planted oil palm can help to shape the ongoing debate about the environmental impacts of oil seed crop expansion, especially if other crops can be mapped to the same level of accuracy. As our model can be regularly rerun as new imagery is published, it can be used to reliably monitor the expansion of the crop in monocultural settings.*"

Please note that we kept the first paragraph of the Discussion in order to present the main results. But we rephrased and restructured the following paragraphs in the Discussion in order to discuss the strengths and caveats in a more logical way: We point out the improvement in terms of accuracy regarding Xu's dataset and our previous study (Descals et al., 2019). Later we discuss the uncertainties associated with the definition of classes based on the comparison of our area estimates and the Gaveau's dataset and FAO statistics. Then, we discuss why we estimated a low ratio of smallholders and the relevance of this finding. Finally, we explain the main shortcoming of CNN: high cost in data labelling and computational resources compared to standard machine learning algorithms.

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).

Response: We have included a new table and figure (Table 1 and Supplementary figure 1), which show the data sources used in the study and the programming environments used in the workflow respectively.

We have also included the URL of data sources in the section *Data Availability*.

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

Response: We have included the reference paper and URL to WorldClim V1 Bioclim dataset in the section *Data Availability*, and also included this source data in Table 1.

The bioclimatic variables were calculated from validated monthly gridded temperature and rainfall records.

The WorldClim V1 Bioclim data was accessed via GEE, but we also provided the link of the official download portal (line 503-505).

Line 107: Ditto for IUCN.

Response: IUCN layer has been included in Table 1. We have made this data publicly available and included the doi in section *Data Availability*.

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.

Response: Landsat images were not used in any processing step. The convolutional neural network classifies Sentinel-1 and Sentinel-2 images, as noted in the submitted manuscript (line 24; line 92; lines 95-96; Figure 1; line 126; line 146; line 269; line 337). The mention of Landsat appears only twice (lines

131 and 276) in order to explain that Landsat was used in previous oil palm mapping studies (IUCN and Gaveau's datasets).

The revisit time in Sentinel-2 (5-days) is higher than Landsat (16 days), which was enough for the generation of cloud-free half-yearly composites globally. We have analysed the amount of empty pixels in the Sentinel-2 composites and found that less than 0.1% of pixels presented missing data. The revisit time in Sentinel-1 is 12 days, but this satellite uses a radar sensor and the presence of clouds in the scenes are not problematic.

The global oil palm layer was generated with Sentinel-1 and 2 composites obtained from images taken in 2019 (line 153, 162, and 496). We did not consider any optimal season with low cloud coverage. For the comparison with the oil palm layer in Xu et al., 2020, we generated a composite for the second half year of 2016, since the latest year in Xu's dataset is 2016. This has been explained in line 279: "*For the comparison with Xu et al., 2020, we used our CNN model to classify Sentinel-1 and Sentinel-2 composites for the second half-year of 2016. [...] we only compared the last year (2016) to ensure data availability in Sentinel-1 and 2 over the study area.*"

We have now included the revisit time in Sentinel-1 (line 151) and Sentinel-2 (line 163), and explain that this revisit time was high enough for obtaining cloud-free composites with Sentinel-2 (line 163).

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?

Response: We have changed the phrasing (line 134): "*Supplementary Table 1 shows the minimum...*".

Lines 132 explained that the statistics in STable 1 were estimated by the authors: "*We estimated the histogram of the nineteen bioclimatic variables in the areas that are classified as industrial oil palm plantations in the IUCN layer.*".

Line 128 explained that the IUCN layer was obtained from another study, in which two co-authors of the current study participated: "*This study's existing oil palm layer was obtained from the IUCN (Meijaard et al., 2018) and shows the industrial oil palm plantations at the global scale.*"

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?

Response: Thank you for noting this. We have now included two additional maps showing the threshold (cut-off) map and the grid cells covering the potential distribution area in Supplementary figure 2. We have explained the cut-off in the caption: "*The upper map shows the number of WorldClim bioclimatic variables that fall within the range observed in the industrial oil palm plantations (IUCN layer). The middle map shows the potential area for oil palm growth, which represents the pixels with more than seventeen bioclimatic variables out of nineteen falling within the range observed in the IUCN layer. The lower map reflects the grid used to cover the study area.*". We have also changed the color shading, which goes from blue (does not meet any bioclimatic variable) to red (meets the 19 bioclimatic variables).

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

Response: We have now corrected the legend (line 140). It now says '*Cells with closed-canopy oil palm detected by the CNN*'.

We also corrected the caption to make clear where closed-canopy oil palm was detected by CNN: '*Cells filled with green show the cells where closed-canopy oil palm was detected by the convolutional neural network (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.

Response: Landsat images were not used in the study, as noted in a previous response. We added information about the central wavelength of band 4 (line 155): "*We also used Band 4 (red band; central wavelength = 665 nm) of Sentinel-2 Level 2A…*".

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?

Response: We used only Band 4 because it shows high contrast between vegetation and bare soil, as explained in lines 157-160: "*Band 4 is the 10-meter resolution band that best shows the roads in industrial plantations because of the high contrast in terms of reflectance between the road and the surrounding oil palm. The high light scattering of vegetation in the near-infrared spectrum makes the recognition of roads less feasible in the 10-meter near-infrared band (Band 8).* ".

NDVI does not provide more information than Band 4 because the near IR (Band 8) saturates not only in vegetated surfaces but also in the surrounding roads.

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?

Response: Indeed, the shapes of the roads are important, as explained in line 195: "In flat surface plantations, the harvesting trails are usually built in straight lines and thus form a rectilinear grid (Figure 3a). In contrast, the industrial plantations that are constructed over steep terrain usually present curvy trails (Figure 3b).". In fact, rather than presence/absence of roads in absolute terms, it is more about the density of harvesting roads, as explained in lines 203: "When smallholder plantations form a large homogenous cluster, this cluster has a less dense trail network than industrial plantations". To the human eye (and ultimately to the CNN), the shape and the density of roads are the predicting variables that discern smallholders to industrial plantations.

Please note that we have now changed the Sentinel-2 images in figure 3 to Sentinel-1 and 2 composites, which better illustrate the differences between industrial and smallholder plantations.

We have removed the mention of mills as contextual information used by the CNN. The mills are easy to recognize to the human eye in the 10-meter Sentinel-1 and Sentinel-2 images. However, the claim that the CNN recognizes the location of mills is too adventurous, particularly considering that mills appear punctually in the plantations.

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

Response: DigitalGlobe images have been also included in Table 1, which presents all geospatial sources used in the study.

We also elaborated the paragraph in which we explain how these high-res images were used: (line 170-174) '*We digitized the oil palm plantations also by interpreting the very-high-resolution DigitalGlobe images that are displayed as the base layer in GEE. The DigitalGlobe images have a sub-meter spatial resolution and are displayed as true-color composites in GEE. These images are updated regularly and the date depends on the location, but usually, the images are taken during the past one to two years. The DigitalGlobe images were used as complementary data to the Sentinel-1 and 2 composites in the visual interpretation.*'

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?

Response: We have developed the paragraph (lines 181-185) in order to better explain how data augmentation improves the classification accuracy and included relevant references that justify this technique. As explained in the response to Comment C (above), we decided not to include details on the quantitative improvement provided by the data augmentation, since this is proven in the cited studies and such analysis seems less relevant to a data journal and more appropriate to a deep learning journal or technical remote sensing journal.

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?

Response: We do not think that the discussion on smallholder versus industrial-scale producers is premature, especially because of the "economic, social, environmental and political significance" of the differentiation between the two forms of production, mentioned by the reviewer. There is ongoing debate in the fields of sustainable agriculture and sustainable development on the optimal scales of production for meeting future oil demand and development goals, and improving the mapping of these different forms of production is important. Nevertheless it is clear that smallholder and industrial oil palm are not uniquely differentiated forms of production but rather extremes of a production scale continuum from individual natural trees in an agroforest setting occasionally harvested for local use to privately owned smallholder farms, smallholder tenants under medium-scale ownership, compulsory smallholder production in industrial oil palm (in Indonesia and Malaysia), to finally large, single-owned company plantings. Differentiating between these forms of production at a global scale is beyond the capability of our current analysis, but this does not mean that we shouldn't highlight that our analysis approximates the differentiation between the two main forms of production and that it allows us to discuss the broad implications of this result.

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 the strengths of CNN and ‘let it run’.

Response: Thank you for pointing this out. We took for granted that ‘CNN can automatically learn contextual information’ owing to the recent advances in the deep learning field. So indeed, we anticipated the strengths of CNN given the current literature about the topic.

We agree with the reviewer that the goodness of the CNN for the classification of oil palm should have been questioned from the beginning. Thus, we have removed the sentence that the reviewer quoted, and we have avoided any anticipation of the good results of the CNN in our case study. However, we have kept the lines in which we mention the good results obtained with deep learning in previous remote sensing studies: (Introduction section, line 82) “convolutional neural networks (CNN) have recently been embraced by the remote sensing community due to the ability to recognize intricate patterns in the images (Ma et al., 2019)”.

The demonstration of the strengths of the CNN was, in fact, one of the goals that we had in mind when we compared the CNN results with previous oil palm datasets. The oil palm maps in Xu et al., 2019 and Descals et al., 2019 were generated with a Random Forest classification. Here, we thank the reviewer again since the suggestion of including an uncertainty analysis has been decisive to prove that the CNN performed significantly better than the Random Forest with a 95% confidence level.

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?

Response: Thank you for pointing this out. During this review process, we also got some inquiries regarding the implementation of the CNN. We have clarified this point in the lines 107-109: “*The visual interpretation of training and validation data was also done in GEE. The training of the CNN and the classification of images, however, was performed with a local computer using Matlab 2019a since the implementation of the CNN model was less feasible in GEE.*”. We have also mentioned the computational resources used in the study: (line 469) “*[...] the computing time for training a pre-trained DeepLabv3+ was nearly 8 days with an office computer.*”

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.

Response: The logic of this paragraph was not clear. We have rephrased this paragraph: (lines 466-472) “*The shortcomings of deep learning include the high computational cost for training the models and the high cost for gathering labeled data compared to the standard machine learning algorithms used in remote sensing, such as Random Forest. In this study, 296 images of 1000 x 1000 pixels were used as a training data set, consisting of 200 labeled images and 96 augmented images, and the computing time for training a pre-trained DeepLabv3+ was nearly 8 days with an office computer. Despite this, the computing time and the size of our training dataset was considerably lower than state-of-the-art deep learning studies in computer vision (i.e. more than 200,000 labeled images in the Common Objects in Context (COCO) dataset) in which the number of classes and complexity of the classification problem surpasses the current study.*”.

Please note that this paragraph aims to introduce a shortcoming of the CNN: the high-computational cost and cost of labelling data compared to standard machine learning algorithms. The main strength of the CNN, which is a higher accuracy than Random Forest, is commented in a previous paragraph.

Reader need more information about DeepLabv3+, especially in a source table. Chen et al. 2018 - obscure conference proceedings - not a useful reference.

Response: We have added the main reference paper to the DeepLab models:

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 834-848.

We kept Chen et al. 2018. This reference, although a conference proceeding, is the only one that introduces the changes in the v3+ of the DeepLab model.

Please note that the code has not been included in Table 1 because we dedicated this table only for geospatial data. We found it more appropriate to include the link to the DeepLabv3+ code in the section *Code availability*.

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.

Response: We have now included an explanation of the sampling design method: (line 257) "*The points were randomly distributed using a simple random sampling, which means that each pixel in the map had an equal chance of being selected.*". We have also added an explanation about why this leads to an imbalanced number of samples per class: (lines 230-233) "*This sample method led to a high imbalance between the points labeled as 'Other land uses' and the points labeled as oil palm, both industrial and smallholder, since oil palm plantations present a rare occurrence in the study area. The rare occurrence of oil palm implied that the probability of randomly selecting an oil palm plantation was also low..*".

Despite the low sample size for smallholdings, the confidence intervals for user's and producer's accuracy were below 7%.

We are aware that we achieved a high overall accuracy partially due to the high amount of points labelled as 'Other land uses'. In fact, if we had classified all pixels as class 'Other', we would get an OA = 92.00 ± 0.51% (this is the so-called no-information rate).

Please note that we have also included the statistical test that shows that the overall accuracy obtained with the CNN is significantly higher than the no-information rate: lines 351-354 "*The accuracy metrics obtained with the 10,816 points show an OA of 98.52 ± 0.20% (Table 2) for the global oil palm map (Supplementary Table 3 shows the confusion matrix). This OA is significantly higher than the no-information rate (92.00 ± 0.51%) and, thus, we can assure that the CNN classification did better than assigning the major class to all the validation points.*".

A description of the statistical test involving the no-information rate has been included in lines 264-269.

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?

Response: Thank you for noting this missing information. We have now specified that the DigitalGlobe images have sub-meter spatial resolution in line171 and also in Table 1.

Lines 171-174 explains now that “*... DigitalGlobe images [...] are displayed as the base layer in GEE. The DigitalGlobe images have a sub-meter spatial resolution and are displayed as true-color composites in GEE. These images are updated regularly and the date depends on the location, but usually, the images are taken during the past one to two years. The DigitalGlobe images were used as complementary data to the Sentinel-1 and 2 composites in the visual interpretation.*”

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.

Response: We have included an explanation of statistics in Section 2.7 *Validation*, added confidence intervals for both accuracies and area estimates throughout the text and figures, included Table 1 with all data sources used in the study, and provided links to all data sources that are openly accessible in the section *Data availability*.

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?

Response: This particular statement was referencing a previous study (Miettinen & Liew, 2011) in which oil palm is detected using only radar. We have now cited also our previous study, in which we used a significance test to prove that the accuracy of the oil palm map improved with the combined used of Sentinel-1 and 2: (line 395-397) *‘The results confirm previous findings on the suitability of radar satellite data for mapping closed-canopy oil palm plantations at the regional scale (Miettinen & Liew, 2011) and the combined used of radar and optical data for mapping smallholder and industrial oil palm plantations (Descals et al., 2019).’*

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?

Response: The sentence might be confusing. Thank you for noticing it. The sentence meant to say that oil palm maps are planned to be generated annually in the future, but using the same CNN model as for year 2019. We rephrased the text to make it clear: (line 465) ‘*The CNN model trained for the year 2019 is planned to be used for follow-up monitoring once a year to generate global oil palm maps.*’

Indeed, there is no reason for retraining the CNN model. The same CNN model used for 2019 will generate oil palm maps in the coming years with the same accuracy as for 2019.

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.’.

Response: The study referred to ‘Gaveau et al. forthcoming’ is under review, but the preprint is now available online (doi: 10.21203/rs.3.rs-143515/v1). We have removed the ‘Gaveau et al. forthcoming’ and cited the preprint instead.

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 quantitatively distinguish industrial from smallholdings. Absent that proof, one doubts that this interesting discussion has any place in a data description.

Response: We have added the confidence intervals in both accuracy metrics and area estimates following the good practices for validating remotely-sensed data by Oloffson et al., 2014.

The main points regarding the goodness and caveats of the oil palm map remain the same (mentioned in answer to Comment D, above).

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.

We have added a description in the landing page of the Github repository, which explains the content of the repository: “*The repository contains the Google Earth Engine (GEE) code for the generation of Sentinel-1 and Sentinel-2 composites. The composites contain three bands: VV band and VH band from Sentinel-1 and band 4 from Sentinel-2. These composites were used as input data in the classification of oil palm plantations at the global scale (https://doi.org/10.5281/zenodo.3884602). The code is included in the file ‘create_S1_and_S2_composites.js’ but can be accessed directly in GEE through the following link: https://code.earthengine.google.com/2e59a630f13906de5f3d0eb696b07ee4*”.

We did not include the update log in the landing page since GitHub automatically generates a report of all changes done in the repository. The changes history are publicly available and accessible in each repository. For the repository in question: https://github.com/adriadescals/oil_palm_global/commits/master

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 ESRIformat, open easily in e.g. QGIS (open source GIS software). .tif files in oil_palm_map files open in most image software.

Thank you for the positive feedback on the accessibility of data.

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.

Response: Following your suggestion (also mentioned in comment A, above), we have included Table 1 which summarizes the data sources used in the study. We have also included the URL link to the sources in the section ‘Data availability’.

The reference to ‘Code availability’ (previously line 417) has been removed from the text. Thanks for noticing this deficient reference. Instead of this reference to ‘Code availability’, the paragraph now includes the link to Sentinel-1 and 2 data via GEE. Additionally, we included the link to the oficial portal for Sentinel-1 and 2 data (https://scihub.copernicus.eu/).

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

Response: We kept the Conclusion as it is since the results obtained after adding the uncertainty analysis and the main points drawn in the Discussion have not changed.

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?

Thank you for noticing this mistake. We were referring to Xu et al. 2020. We have corrected the header label accordingly.