Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-285-RC1, 2022

Comment on essd-2021-285

Anonymous Referee #1

Referee comment on "GISD30: global 30□m impervious-surface dynamic dataset from 1985 to 2020 using time-series Landsat imagery on the Google Earth Engine platform" by Xiao Zhang et al., Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2021-285-RC1, 2022

Comments to manuscript

This manuscript was trying to derive a new time-series (every five-year interval from 1985 to 2020) impervious surface dataset from Landsat imagery with the aid of the Google Earth Engine (GEE) platform. The authors divided the land surface into 961 5°×5° geographical tiles, and used random forest classifiers to identify impervious surfaces in each tile and period. Then they adopt a temporal consistency correction method to smooth the independent results classified during different five-year periods. A satisfactory performance was claimed (an overall accuracy of 91.5% and a kappa coefficient of 0.866) using more than 18 thousand validation samples. However, there are a number of concerns on the framing and introduction, the scientific contribution to literature and current datasets, the clarity of the methodology and results, and most importantly, the validation of the derived dataset. Below I highlight these key areas.

The frame of the introduction

In the introduction, the authors reviewed many previous studies and existing impervious surface products, and summarized a series of existing problems, such as significant inconsistency and uncertainty within existing datasets (L65-69), the monitoring efficiency of the time-series change detection strategy being very low (L81-82, not necessarily true), image classification strategy performing well but collecting training samples being time-consuming and labor-intensive (L94-95). However, all these problems raised are not solved in the current version of manuscript and the objective of this study is missing. Are the authors going to reduce the uncertainty in inconsistent areas, or trying to improve the efficiency of the classification procedure? There have been already serval impervious surface products available, and why the authors still tried to derived a new one? The text in the last paragraph of the introduction "The aim of the study was to automatically produce an accurate and novel global 30 m impervious surface dynamic dataset (GISD30) for 1985 to 2020…" (L102-104) is more like a task rather than a study to solve particular scientific problem(s). I suggest that in the introduction the authors need to focus more on the what is/are the most urgent problem(s) in current studies, what are the main challenges behind these problems, and why the proposed methods / strategies are capable of solving these problems.

The methodology

1. As described in L135, the authors extracted the impervious surfaces in 2020 from the GLC_FCS30-2020 (Zhang, et al., 2021) and used it as a baseline to derive the time-series impervious surfaces from 1985 to 2020. It is not clear what is the difference between this extracted impervious surface layer and the impervious surfaces derived by the authors. Is

the extracted impervious surface layer directly used as the result for period of 2015-2020?
2. The authors mentioned plenty times of their previous studies in the methodology section, such as Zhang et al., 2018, Zhang et al., 2019, Zhang et al., 2020, and Zhang et al., 2021. I am aware that the method used in this manuscript was developed based upon their previous ones. But most of these descriptions should be moved to the introduction and leave only the most original ones in the methodology section. Again, many discussions on other previous studies are found in the methodology section, too. They should also be moved to the introduction. In the current version, I can hardly see the core method proposed by the author or the adaptations/modifications of previous methods to derive time-series impervious surface.
3. Perhaps I missed it but I couldn't see how many training samples the authors used to calibrate the random forest classifier in each 5°×5° tile and five-year interval.
4. According to the manuscript, the global training samples were automatically collected from the GLC_FCS30-2020 land-cover product. Did the authors check the reliability of the labels in these training samples? Even though a high accuracy was achieved in this GLC_FCS30-2020 product, false labels still exist when the training samples are located in incorrect detected areas (impervious surfaces, cropland, bare land, and other pervious surfaces). These training samples with false label can directly bias the classifiers which were later on used to identify the time-series impervious surfaces.
5. Besides, the equation in L264 is unnumbered, and it is the only equation throughout the manuscript. I believe this is not rigorous enough for a manuscript considering for publication in ESSD. I suggest the authors add more equations to better describe the techniques used in this study.
6. The authors said that they independently train the classification models at each period (L311-312). How did they collect the training samples and obtain the labels in period other than 2020?
7. How to deal with the situation where there is no Landsat imagery available? The author presented the availability of Landsat observation during each five-year interval (figure 1). However, the proposed method uses seasonally composited imageries as input features of the classifier (L243-247), and there may not be cloud-free imagery available especially for the rainy season before 2000 in the tropical rainforest areas.
8. The authors did not mention the exact time-span of Landsat imagery used to derive each impervious surface layer. I guess the time-spans correspond to the periods presented in figure 1. For example, they used Landsat imageries during 1991-1995 to derive the impervious surface layer for 1995. But I believe it is more reasonable to use imageries before and after the target year, for example, using imageries during 1993-1997 to derive the impervious surface layer for 1995.
The results
1. The spatial distribution of the time-series global impervious surface presented in figure 5 does not provide much spatial details nor temporal dynamics for the impervious surface because of relatively small size of impervious surface expansion compared to the entire terrestrial surface. Local enlargements of hotspot area should be presented here for better illustration of the results.
2. Again, the comparison between the derived pattern and other available products in figure 10 is not clear from the global view. Local enlargements should be presented, too.
3. There are several problems/confusions in the scatter plots in figure 11. The authors did not give clear labels to the axes. I guess they refers to the proportions of impervious surface after aggregation into the coarse resolution. The color map below the scatter plots is not clear, too. Maybe the color refers to the scatter density in the plots. The root-mean-square error (RMSE) and the coefficient of determination (R2) presented here are mathematically/statistically incorrect, because these two indicators are usually used as a measure of how well the reference data (observed outcomes) are replicated by the model outputs, while the compared products here are not ground truths. Instead, the authors should present the measure of correlation coefficients. I don't know the exact geographical distributions of the scatters, but I believe the numerical distribution of the

scatters is problematic. There are too many scatters located at/closed to 0-value, while only a small proportion of them fall within the range of 20%-100%. This will largely bias the slope and intercept of the fitting lines. The authors should reduce the number of 0-value scatters and at the same time increase the number of scatters with higher values (do not include too scatters with 100% value).
4. The paragraph related to figure 11 (L509-520) should be rephrased to explain why the derived results yields smaller proportions of impervious surface compared with all other pervious products. Are the derived results underestimate the actual situation? Moreover, it confuses me that the results in figure 11 are somewhat contradictory to the results in figure 9. According to results in figure 9, the areas of the derived impervious surface are larger than those of the GAIA, NUACI, and GHSL across different continents. But the results in figure 11 show smaller proportions of derived impervious surface.
5. To justified the outperformance of the derived dataset, the authors should directly compare the accuracies of the derived and previous datasets using the same validation samples, since the accuracies (OA and kappa coefficient) can vary greatly when calculated using different validation samples.
6. The author should include the recent GAUD product (Liu et al., 2020, cited by the authors) into their comparison, since it provides annual impervious surface layers.
7. The comparisons presented in figure 12, 13 and 14 are misleading for product GlobeLand30. Many impervious surfaces existed before 1995 are colored with yellow (2000-2005), which is obvious incorrect.
Reliability of the assessment
1. According to the manuscript, the locations of the validation points were randomly generated using the stratified random sampling strategy (L150). In order to better evaluate the results, especially the time-series dynamics of the impervious surface, I suggest a substantial number of validation samples should be placed within (impervious surfaces) and around the fringe of the urban areas (pervious surfaces), which are exactly the most inconsistent and uncertain areas regarding the impervious surface classification problem.
2. Eyeballing the map of figure 2 there seems to be considerable number of samples located far away from the urban areas. These areas are well detected as pervious surface by many previous methods, or can be easily masked out for example by nightlight observations. Validation samples within these areas do not contribute to justifying the superiority of the derived results, but instead just smoothing the performance difference between the derived and previous datasets.
3. The number of validation samples in each five-year interval should be large enough to evaluate the impervious surface dynamics during the corresponding time periods. According to the manuscript, the land surface is divided totally 961 geographical tiles (5°×5°) and only 18,540 validation samples were used (L150). I did a rough calculation. There are only ~20 samples on average within each tile, which is definitely too sparse for the evaluation of 30 m classification results in a large area of 5°×5° tile, not to mention that these ~20 samples were divided into eight time periods. Great uncertainties/bias are expected in the calculated OAs and Kappa coefficients with these samples.
4. As presented in table 1, only hundreds of validation samples were used to evaluate the performances of each five-year period. If I read it correctly, there are less than 1 sample on average within each 5°×5° tile for each five-year period. Assessment with these validation samples is definitely unreliable and cannot truly reflect the quality of the derived dataset. The authors should substantially increase their number of validation samples to achieve a more reliable assessment.
5. Apart from point-based validation, the area-based validation strategy, i.e., visual interpretating impervious area in small blocks near the urban fringe and comparing them with the derived results, is more encouraged, considering the sparse distribution of impervious surfaces
Uncertainties
The uncertainties of the derived impervious surface layers should be discussed but missing in the current manuscript. This dataset at least involved uncertainties from four aspects:

1) the labels of training samples directly collected from the GLC_FCS30-2020 rather than visual interpretation. 2) the migration of reflectance spectra of impervious surfaces measured in 2020 to other periods. 3) the reliability of the random forest classifiers for each tile and time period. 4) the temporal consistency correction used to smooth the independent classification results. It is not clear how these uncertainties propagate along the entire derivation procedure and to what extent these uncertainties contribute to the final derived dataset.

Others

1. There are many typos in the manuscript. To name only a few for example from the methodology section: brackets in L220, L223, L239, L271.
2. Some paragraphs only consist one or two sentences, e.g., L178-183, L191-196, L214-218.
3. Although the manuscript is readable, there are still many inappropriate expressions and grammar errors throughout the entire manuscript. Please consult a native English speaker or a commercial proofreading service.