Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-278-RC2, 2022

Comment on essd-2021-278

Anonymous Referee #2

Referee comment on "A high-resolution inland surface water body dataset for the tundra and boreal forests of North America" by Yijie Sui et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-278-RC2, 2022

General comments:

This manuscript presents a data set of inland surface water bodies (SWBI) for the tundra and boreal forests of North America. The data inventory is generated by an automated approach of mapping the 10 m resolution Sentinel-2 multispectral satellite imagery. The resulting SWBI includes approximately 6.7 million water bodies > 0.001 km^2, of which there are 6 million (~90%) smaller than 0.1 km^2. The data set is also compared with other earlier regional or global water body products (e.g., JRC GSW, PeRL, GLWD, and HydroLAKES) and manually interpreted data. The data set, if with good quality, can provide finer-scale water body distribution information over the tundra and boreal forest regions of NA Arctic, which is critical for studying Arctic surface water bodies in response to changing climate and thawing permafrost. However, several fundamental problems related to the remote sensing mapping method and insufficient data quality check prevent the paper from being considered for publication.

1. Mapping method

(1) The water body mapping method in section 4.1 is not described clearly. The logic of mapping procedures, including mapping updated water frequency, applying the machine learning model for water body identification, and deriving the final water body map is not described sufficiently, confusing the audience about the rationality of the methodology.

(2) The mapping method integrates the Sentinel-2 derived water frequency and JRC water dataset. However, the JRC dataset generated from Landsat imagery has a spatial resolution of 30m, which is inconsistent with 10-m Sentinel-2 data. How to create a 10-m SWBI data set with the water body size as small as 0.001 km^2, about one pixel for Landsat imagery? Another water frequency data product based on Landsat imagery, the GLAD group's Global surface water dynamics 1999-2020 (https://geog.umd.edu/news/new-global-surface-water-dynamics-maps-published-remote-sensing-environment), may be also considered for the integrated mapping.

(3) The mapping algorithm (Eq. 4) adopts a weighted linear combination. The weights and water frequency thresholds seem to be determined arbitrarily. In addition, whether the training set size (1250 points) is sufficient to establish the machine learning models for mapping approximately 6.7 million water bodies requires more evidence.

2. Data quality control

The reviewer roughly went over the SWBI dataset. Many mapping errors exist in the data product, e.g., mixing with ocean water areas near the coastline, remaining river segments, the rough and tumble polylines, etc. The flowchart (Figure 3) suggests that an identification procedure was conducted, excluding rivers/streams and removing noises. However, many river segments and coastal waters still exist in the data set. How to separate the multiple lakes linked by river/stream? This will influence the number and morphologic calculation of water bodies in the data product. I suggest severe quality assurance before publishing the data set.

3. Data validation

In this study, the data quality assessment includes two aspects, the comparison with other data products, e.g., JRC GSW data, HydroLAKES, and GLWD, and the validation by manual interpreted data. However, the earlier data products have inconsistent water body definitions with the SWBI. For example, the raster-format JRC GSW contains all water bodies, e.g., lake, river/stream, fish ponds, etc., while the HydroLAKES and GLWD only include lakes (and reservoirs). It is therefore not valid to attribute the differences to mapping quality.

Other comments:

Line18: This data set does not include water bodies in Eurasia Arctic. How to say it is a more complete representation than the PeRL data?

Line19: I would like to recommend the public share and open access of the manually interpreted data, which can be helpful for the quality assessment of the future published water body data sets.

Line82: "Lakes and ponds" used here have different means?

Line83: "…about 50% of the lakes and 30% of lakes by area…" should be "…about 50% of the lakes by count and 30% of lakes by area…"?

Line97-98: For the Sentinel-2 sensors, there are three bands at the 60-m resolution among the total of 12 bands.

Line131: The MNDWI calculation requires the input of SWIR band, with a resolution of 20m for Sentinel-2 imagery. How to process the different resolutions for different spectral bands?

Line141-143: The references for the NDWI (McFeeters, 1996) and MNDWI (Xu, 2006) were mistakenly cited.

Line155: how to determine the threshold of "hue <0.45" for extracting water pixels? Please test the threshold sensitivity for water bodies under different conditions and images.

Line161: A is the updated water frequency. What is the final mapping result of water body data set?

Line162: how to derive the Sentinel-2-derived water frequency (As) here? By the machine learning model as introduced in the following part?

Line163: the threshold setting by combining water frequency and elevation looks a little weird. What is the rationality for doing this?

Line167: "(Figure 4a)"---the wrong inserting place.

Line190-191: the terrain shadows have few influences on the water body mapping for the areas with an elevation below 1500 m?

Line197-200: As shown in the data set, the polygons of mapped water bodies were generalized by GIS tool. Does the simplification tolerance affect the calculation of geometry metrics of polygons?

Line243-246: The analyses of water body abundance (the power-law statistics), the counts of different lake size levels for the SWBI, should be added.

Figure 7a: the 5 km × 5 km grid can contain the water area >500 km^2???

Line262-263: please indicate the sub-title for maps (a, b, c, and d).

Line273-280: why are the mapped water body areas mostly smaller than the manually interpreted data?

Line315-316: the number of SWBI water bodies for the size levels (100~1000, 10~100, 1~10, 0.1~1) are all slightly smaller than that of HydroLAKES, why? for lake changes (disappearances) during different mapping periods, or mapping uncertainty?