Reviewing process:

Chen, X., Mu, C., Jia, L., Li, Z., Fan, C., Mu, M., Peng, X., and Wu, X.: High-resolution dataset of thermokarst lakes on the Qinghai-Tibetan Plateau, Earth Syst. Sci. Data Discuss. [preprint], https://doi.org/10.5194/essd-2020-378, in review, 2021.

General comments:

The manuscript by Chen et al 2021 tackle an important global hiatus regarding the lacking geographical delineation and monitoring of small waterbodies (< 10 000 m^2) in the Earth's surface. The authors focus on the importance of having a better inventory of thermokarst lakes, which are known to be biogeochemical important. The type of research developed by Chen et al 2021 is crucial to support with more complete scientific data to feed Earth System Models and Global Climate Models. To develop the dataset, the authors use Sentinel-2 satellite imagery, taking advantage of its spatial resolution and totally free to use policy. The methodology is not innovative, since it was based on the thresholding definition of a very well know and widely used water index (Normalized Difference Water Index – NDWI) and cloud computing based on Google Earth Engine (GEE), although the results are very interesting when cross-validating with very high resolution satellite imagery, for instance, in Google Earth Pro. The georeferentiation quality seems also good. These results clearly highlight the importance of using Unmanned Aerial Vehicles (UAV) field surveys for better estimate the errors and also the importance of performing intensive visual interpretation of the satellite imagery in order to remove outliers and guarantee the best results. The three months of intensive visual interpretation of more than 100 Sentinel-2 scenes made by the authors is remarkable and was worth it. The simple use of the NDWI for mapping the lakes, using a unique threshold was possible in the Qinghai-Tibetan Plateau (QTP), due to its homogeneous landcover characteristics, but the methodology is not replicable for other permafrost areas, for instance in the boreal forest/tundra zones, were landcover is known to be very heterogeneous. The authors support their work in the work of other authors, using their datasets and trying to provide new advances in the formation and development of thermokarst lakes in the QTP, although the definition of thermokarst lakes that the authors use is very broad and it is probable that some of the lakes that the dataset contains did not have a thermokarst genesis. Field work is still lacking regarding this aspect. To conclude, the dataset created and freely available in a widely used format (shapefile) can be very useful in order to support further methodologies and research, concerning, for instance, small lake expansion/drainage events and the spatial and temporal dynamics of some of its optically active constituents that are known to correlate well with hyperspectral and multispectral data, such as Dissolved Organic Carbon (DOC), Total Suspended Solids (TSS) and others, and with this provide a better understanding on their biogeochemistry role in the thermokarst landscapes of the QTP.

Specific comments:

1. Line 24: The authors write: "the true spatial distribution by using a resolution of 10 m with a relative error of 0-0.5", however this a very strong statement. In addition, this relative error, in the way it is, is not easy to understand. How this error was assessed is not clear in the manuscript.
2. Line 38: I suggest eliminating the adjective: "obvious".
3. Line 65: I suggest eliminating the adjective "obvious" to characterize the permafrost degradation of the QTP. Many adjectives in the introduction should be replaced by numerical and quantitative information for a consistent scientific writing.

4. Line 113: Why the authors only mention Sentinel-2A data? Did not the authors use Sentinel-2B data? Why? This is not clear. The authors mention a revisit time for Sentinel-2A of 10 days, but then talk about the twin-satellite system (Sentinel-2A and Sentinel-2B) that have a revisit time of 5 days, without using Sentinel-2B data?
5. Line 122: This sentence is confusing. The Sentinel-2 data is from the European Space Agency (ESA). Note that the Earth Explorer from the United States Geological Survey (USGS) is only an intermediate service to download and access the data. Did not the data was downloaded and accessed in Google Earth Engine? Even in this platform the data provider is the European Union/ESA/Copernicus.
6. Line 135 (figure 1): I would suggest some context information like place labels and also the country boundaries delineation in the small scale map in the upper left corner. Was the map made by the authors or is it from Zou et al (2017)? The subtitle has to be more complete, perhaps mentioning all the sources of information where the auxiliary data came from.
7. Line 138: Why did the authors choose the SRTM of 90 meters instead of the one of 30 meters?
8. Line 140: The end of the sentence is confusing. Which interpolation method are the authors talking about? There are a lot of interpolation methods and this seems a small step of the entire process of generating the SRTM to be highlighted.
9. Line 148: It is not clear why this division was made.
10. Line 172: The authors should support the first sentence of this paragraph with references. See for example: Bouchard, F., MacDonald, L. A., Turner, K. W., Thienpont, J. R., Medeiros, A. S., Biskaborn, B. K., Korosi, J., Hall, R. I., Pienitz, R., & Wolfe, B. B. (2017). Paleolimnology of thermokarst lakes: a window into permafrost landscape evolution. Arctic Science, 3(2), 91–117. https://doi.org/10.1139/as-2016-0022.
11. Line 183: The GEE not only provide MODIS and Sentinel satellite data, but from other satellite constellations also (e.g. Landsat, Sentinel-5 and more). Please, clarify this. The next sentence (beginning at the line 185) has also to be reformulated. Some words are missing.
12. Line 189: Why did the authors use Sentinel-2A L1C data, instead of L2A? Were the images atmospherically corrected and compensated? Did the authors apply the Bidirectional Reflectance Distribution Function (BRDF) and took as reference a Digital Terrain Model to eliminate topographic shadows, for instance in mountain areas? How did the authors manage to solve this type of problems? When visual inspecting the authors database I was able to find some lakes in the mountain areas that were in fact mountain shadows (e.g. Northwest of Tarim). The authors are asked to gently eliminate this type of features and artifacts from the database.
13. Line 192: What do the authors mean with “green light band”? The authors should use references to demonstrate NDWI effectiveness and better supporting its theoretical and physical basis.
14. Line 194: The authors mention the resolution of Sentinel-2 bands previously. This, in the way it is, seems a repetition. Plus, the SWIR bands are not used in this index.
15. Line 215: Why did not the authors use the FMask algorithm (or similar) to previously remove some of this noise of the images, such as clouds, snow and clouds shadow?
16. Line 232: Was the NDVI data extracted from MODIS or Landsat 8? This is not consistent with what the authors mention at the line 136 and further in figure 2.
17. Line 239: Which UAV was used? How was the UAV data processed? Which technique was used? What were the Root-Mean-Square (RSM – X, Y and possibly Z) errors of the generated orthomosacis? These are important details to mention, specially when studying waterbodies due to the lack of contrasting features that make unfeasible the use of some techniques, such as Structure from Motion (SfM). The authors are gently asked to provide new advances on this topic.
18. Line 251 (table 1): How was the relative error calculated? What does it mean? Does the error have units? This is not clear throughout the entire manuscript.

19. Line 260 (figure 4): Although the distribution of thermokarst lakes in the QTP is very interesting, in this map it is not possible to discriminate and understand that distribution. I suggest the authors using a generalization procedure or less classes in order to highlight better all the features (e.g. polygon to point just for visualization purposes in this map or other type of generalization).
20. Line 369: The authors are asked to add a reference in this last statement since some authors also have demonstrated this. See for example: Turetsky, M. R., Abbott, B. W., Jones, M. C., Anthony, K. W., Olefeldt, D., Schuur, E. A. G., Grosse, G., Kuhry, P., Hugelius, G., Koven, C., Lawrence, D. M., Gibson, C., Sannel, A. B. K., & McGuire, A. D. (2020). Carbon release through abrupt permafrost thaw. Nature Geoscience, 13(2), 138–143. https://doi.org/10.1038/s41561-019-0526-0.

Technical corrections:

- Sometimes the English is confusing to the reader and I suggest the authors to fully rewrite some of the sentences and ensure that they are clear for the reader.
- The authors used to write as units m^2km^2 in many circumstances, but this wrong. The authors should uniformize the working units and fix these issues throughout the entire manuscript.
- If the authors choose to use acronyms, they should use it all the way throughout the manuscript.
- Line 169 (figure 2): Sentinel-2A is not well written in both situations. Since the authors add the source of information in the first row of the figure, I would suggest adding the SRTM right above ALT, and change this last name for topography to be easier to understand.
- Line 220: Online waterbody extraction? What does this mean?

Thank you.