Referee #1

General Comments

Mullen and Muller present a new method for producing time series of water extent in large, rapidly-changing and ecologically/culturally/economically important lakes. They use a novel approach implemented in Google Earth Engine (GEE) and validate their results against existing historical data, finding their method to work well, except when scenes contain snow/ice. Overall, the method is robust and the writing and figures in the manuscript are generally clear. However, I have several major concerns with the paper, chiefly related to the discussion of the method's limitations and the situation of this paper within the broader literature, described below. There are also several typos, missing commas/parentheses and some incomplete sentences in the manuscript. I am not certain I caught all the errors, so I suggest the authors carefully edit the paper again before submitting a revised version.

Response

We thank the reviewer for their thorough and helpful review. We interpret the reviewers' comments as requiring (i) a more thorough discussion of the specific contribution of this manuscript in the context of the large literature on remotely sensed water detection, and (ii) a more careful analysis on the limitations of the approach. We address the reviewer's first point with a substantial rewrite of the introduction where we cast the proposed approach as addressing a gap filling challenge that is specific (and essential) to the type of high frequency historic reconstruction that we seek to achieve. We address the reviewer's second point by adding a set of numerical experiments exploring the sensitivity of the approach to different types of classification errors in the input data (i.e. clouds mistakenly classified as water or land, or water and land mistakenly classified as clouds), which we believe are the main limitation of our approach. We also added three lakes in the validation analysis to illustrate the application of the method in a small (~1km2) lake and in a mountainous/snowy setting.

We hope these new elements in the revised manuscript (specified in **red** below), and our specific responses (in **blue**) to their comments, adequately address the reviewer's concerns.

Major Comments:

1. I would have thought that the specific cloud masking method could have a significant effect on the results, yet the cloud masking is only described in the SI and not given much attention in the manuscript. More discussion of the cloud masking method is needed in the main text. Furthermore, I would also suggest additional analysis and discussion about how the choice of a certain cloud-masking algorithm may or may not affect the results. For example, questions that I feel need to be addressed include what percent of pixels are cloudy/poor quality? How does this vary by lake/by year? Do lakes with greater cloudiness exhibit higher error than lakes with lower cloudiness?

Response

We agree with the reviewer that cloud detection is an important prerequisite to our approach that merits further investigation. We will add a discussion section focusing on this dependence in the new manuscript, along with the following analyses:

- For each lake where in-situ validation data is available, we will add a scatterplot with the prediction error on the lake area plotted against the percentage of cloud cover. Results show that, across lakes, the prediction error does not increase significantly with cloudliness for the particular cloud masking algorithm that we use.

- We use numerical experiments to explore the sensitivity of our approach to the accuracy of cloud detection more generally. To investigate the effect of an overly eager cloud detection that tends to overestimate cloud cover, we changed a number of randomly selected (land and water) pixels of each image to cloud, and evaluate the method's ability to determine their original class (water or land). To investigate the effect of an algorithm that fails to fully capture (i.e., underestimates) clouds or cloud shadows, we "flipped" a number of randomly selected pixels from water to land (and vice versa) between the supervised and unsupervised classification steps of our algorithm. This emulates the fact that undetected cloud pixels might be misclassified as land or water. We then assess the effect of this misclassification on the method's ability to predict the class of clouded pixels. We find that our approach is robust to the former, but sensitive to the latter, type of error. The approach is therefore more compatible with an overly eager cloud detection algorithm, rather than an overly greedy one.

Text modifications

- Added scatterplots of errors vs. cloudliness for the validation lakes in Fig 3. These will replace the scatterplots currently in the middle column of the figure which are redundant with the reported R2 statistic.
- Added Figure and result section describing the numerical experiment.
- An added discussion section on the effect of "too little" information due to cloudiness, and overly eager cloud detection method or a smaller lake size.
- Modified conclusion section focusing on the practical implications of the method. We will specifically discuss the effect of an overly eager or greedy cloud detection algorithm.

2. The authors test their method over a small number of lakes – only 6 in total. But given the global availability of Landsat data, and the plethora of studies examining regional to-global scale variability in surface water extent using Landsat/GEE (see comment #3), analyzing over only 6 lakes seems to me like a very small sample size. I encourage the authors to consider adding additional lakes to the analysis, perhaps with different environmental conditions such as in areas with high topography/high latitude (see comment #4).

Response

While we agree with the reviewer that a larger sample size would be ideal, a persistent challenge that we ran into was to find reliable in situ observations of lake extents that matches

the monthly frequency and multi-decadal observation periods that we are targeting. Oftentimes we would find lake elevation time series but no reliable elevation-area relationships to obtain lake extents. Short of having a large sample of validation lakes, the revised manuscript uses a combination of targeted case studies and numerical experiments to investigate key limitations of our approach. Specifically, in response to comment #3 below, we now emphasize that our approach serves as a gap-filling algorithm, which purpose is to determine the class (water or land) of masked (cloud) pixels. We hypothesize that two main challenges can affect the accuracy of the gap-filling method: (i) too little information is provided in the input imagery or (ii) wrong information is contained in the input imagery or inundation frequency raster.

In the revised manuscript, we investigate these two limitations (too little information and wrong information) in two ways:

1. We use the two numerical experiments described above to investigate the effect of (i) too little (randomly masked pixels) and (ii) wrong (randomly flipped pixels) information. We add a third experiment where we add random noise to the inundation frequency raster. Comparing the outcome of the two latter experiments allows to differentiate the effect of having wrong information in the *dependent* (randomly flipped pixels) or in the *independent* (noise in the inundation frequency raster) data used to train the supervised classification.

2. We add three validation lakes to illustrate the effect of each limitation. Two small (~1km) lakes in Texas illustrate the effect of having too little information (i.e. a small number of available pixels). The other lake in upstate New York illustrates the effect of having wrong information, as the cold climate and mountainous terrain introduce errors in the unsupervised classification of clouds, water and land.

Both analyses suggest that the proposed method is more directly limited by wrong information, rather than too little information. This is consistent with the discussion on cloud masking of the previous comment. An overly eager cloud detection algorithm will err in favor of providing too little (rather than wrong) information and have a smaller effect on prediction performance.

Text modifications

- Three validation lakes added to Figure 3:
- An added figure and result section with the three proposed numerical experiments.
- An added discussion section on the effect of “too little” information due to cloudiness, and overly eager cloud detection method or a smaller lake size.
- We will modify the current discussion section focusing on water detection challenges to discuss the numerical experiment and focus on the drivers and effects of “wrong” information in the input classified imagery.

4b Relatedly, the authors should also consider adding discussion about the implementation of the method and the ease of running it – i.e. is the method computationally slow and therefore would be challenging to run over large areas or could this be reasonably run over, say, hundreds of large reservoirs?

Response

With regards to implementation and scaling, we foresee no theoretical limit to the scalability of the proposed gap filling approach, thanks to Earth Engine's massive parallelization capability. The only "hard" limit lies in the need for users to define a region of interest (ROI) over which to carry out the supervised classification. In the manuscript, these ROIs were determined manually by roughly following the maximum footprint of the considered lake. Methods can be developed to generate regions of interest automatically, for instance by creating a buffer around a certain threshold in our inundation frequency raster. However, we feel that developing and validating these approaches goes beyond the scope of this paper. We added a short discussion in the conclusion summarizing the above points. We also link to a working Earth Engine script for readers to evaluate for themselves the practicality of implementing the method.

Text modification

- Modified conclusion section focusing on the practical implications of the method. We will specifically discuss potential and limitations from a computational implementation perspective.

3. This manuscript requires additional discussion of how this method fits in with the (very large) literature on monitoring lake extent using optical satellite imagery. The manuscript makes little mention of the work of Pekel et al. Nature, (2016), who map global variability in water extent using Landsat and GEE, or regional studies such as Zou et al. PNAS, (2018) or Wang et al., Nature Geoscience, (2018), or even the large literature on reservoir monitoring using MODIS or other optical sensors (e.g. Gao et al., Water Resources Research, 2012). While I do appreciate that this method is designed to produce highly accurate time series for individual lakes which is different than the goals of many of these other studies, I feel more discussion is needed to distinguish specifically how this method is an advance compared to this previous work and particularly, what specific scientific questions this approach could answer that other approaches could not.

Response

The reviewer raises a good point and we edited the introduction to better describe the scope of the paper and its place with respect to previous work. In particular, we now reframe the general challenge of reconstructing historical time series of lake extent as a two-step problem. Only the second step is addressed by the proposed approach, although its sensitivity to errors from the first step are fully investigated in the revised manuscript (see comments 1 and 2 above).

1. The first step concerns the detection of land, water and clouds using multispectral imagery. This is a well studied problem that is generally well addressed, although well-known issues arise under specific circumstances. Improving the detection water, cloud and land from a

pixel's spectral signature is beyond the scope of this paper and we refer to the appropriate literature in the revised manuscript.

2. The second step of the problem is in essence a gap-filling challenge, where pixels classified as clouds during the first step must be reclassified as water or land. This problem is relatively easy to address if images are available at a short return time (e.g., daily MODIS imagery). For example, monthly water cover can be obtained from daily MODIS images, by simply masking clouded pixels from the analysis and taking the per-pixel mean inundation status of each stack of overlapping unmasked pixels. Similar approaches are used in several of the global studies mentioned by the reviewer (a detailed review of which is provided in the revised manuscript), where the main challenge lies in the classification of water, cloud and land (step 1), rather than the inference of the flooding status of cloudy pixels (step 2). Unfortunately, a similar approach cannot be applied to Landsat images due to their longer (two weeks) return time. Although Pekel's(2016) dataset is unique by providing monthly high resolution global water cover grids going back to the 1980's based on Landsat imagery, masked (cloudy) pixels are left unclassified and indistinguishable from missing pixels (e.g., due to Landsat 7 sensor failure or missing images). This leads to a significant underestimation of lake water extents (see Zhao and Gao 2018, GRL). The approach proposed in the manuscript seeks to address this type of situation and infer the classification status of mask landsat pixels. Its novelty, compared to previous work addressing the same issue, is its unique use of supervised classification to leverage historic inundation frequency (see response to comment #1 from reviewer 2).

Text modification:
Rewritten introduction to reflect the points made in our response above

4. Relatedly, I also feel this manuscript is lacking some discussion about limitations and specific applications. The discussion about the different assumptions of the method is good; however, I was left wondering more specifically where this method might work and where it might fail. For example, would this method work in areas of high topography/high latitudes where topographic shadowing is an issue? What is the smallest lake this method would work on? Is there a relationship between cloudiness/size/error? I would also advise more discussion about what might have caused the outlier points removed in the time series analysis.

Response
We thank the reviewer for their comment and hope that the new discussion section about the approach's sensitivity to classification errors in the input imagery (comments 1-2 above) will address their concerns. Many of the questions asked by the reviewer in their comment boil down to the effect of pixel misclassification (or, more fundamentally, to too little or wrong information) on the method's ability to re-classify as land or water the pixels that were previously identified as clouds. The new analyses in the revised manuscript (comments 1-2 above) show that the method is robust to the former (too little information) but sensitive to the latter (wrong information).

In practice, this means that the method will not perform well in locations where circumstances (topographic shading, cloud shading, snow/ice, etc) makes it difficult to reliably distinguish water from clouds and land using multispectral imagery. The inability to properly classify clouds, water and land in the input imagery gives rise to the outliers that are subsequently removed from the time series analysis. This point is clarified in the revised manuscript when discussing the numerical experiments and the added lake from upstate New York.

In contrast, the method is likely to be generally robust to application to smaller lake sizes, despite a limited number of unmasked pixels available to train the supervised classification. However, performance is contingent on the assumptions that (i) the location of cloud and water pixels are statistically independent and (ii) that lake bathymetry does not change over time. We posit in the original manuscript that these assumptions are less likely to be satisfied in small lakes, but in situ data is insufficient to formally test this hypothesis. The two assumptions appear to be satisfied for the ~1km2 lakes that we added to the validation sample in the revised manuscript.

Text modification:

- Added figure and results section on the numerical experiments
- Added discussion section on the effect of "too little" information due to cloudiness, and overly eager cloud detection method or a smaller lake size.
- We will modify the current discussion section focusing on water detection challenges to discuss the numerical experiment and focus on the drivers and effects of "wrong" information in the input classified imagery.
- Modified conclusion section focusing on the practical implications of the method. We will specifically discuss accurate water detection in unmasked pixels as a key prerequisite to the method.

Specific Comments:

Response

We thank the reviewer for their detailed specific comments, which we think generally improve the readability of the manuscript. We will address all comments in the revised manuscript, provided that they still apply (i.e. if the referred text was not already modified to address a major comment).

L1: "The empirical attribution of 'past' rapid hydrologic change" L15: change "when applicable" to "where available"

L15: I would advise adding a sentence at the end of the abstract stating the importance/broader significance of your findings, instead of just stating that your method works

L18: In my opinion, the first few sentences of the paper are weak. I would suggest rewriting slightly (i.e. "Despite their importance, many lakes are undergoing rapid change. . ." makes little sense – the importance of lakes doesn't necessarily mean that lakes will not or should not

undergo rapid change). Since “rapidly changing” is a key part of the manuscript, I would also suggest defining what you mean by rapid change paper since the time scale implied by “rapid” can vary based on the reader’s background.

L27: This sentence (“By providing”) should start the next paragraph, not sit at the end of this one as it interrupts the flow

L31: The paragraph starts by talking about monitoring surface water extent, but then discusses radar altimeters before moving back to extent. I would suggest restructuring this paragraph, or at least the first sentence of it, as the current structure is confusing

L83: I suggest adding a sentence or two to the final paragraph of the introduction stating something like “we test this method over XX lakes, analyze its accuracy and demonstrate its utility” just to provide readers with a better road map for the manuscript

L86: I would call this first step something like “Masking” instead of pre-processing (see major comment #1 above).

Figure 1: I like this figure, but think it could be improved slightly by increasing the size of the image panels and decreasing the size of the arrows and white space. The image panels are hard to see in places and there’s plenty of white space so it should be straightforward to make them a bit larger and easier to see.

L149: The sentence starting with “Indeed, visual inspection of satellite. . .” is unclear. I think what the authors are stating is that the area-elevation curves do not match the satellite-observed area, but this section could be clarified.

Figure 3: Please make the x and y labels and the symbols themselves much larger, it is nearly impossible to read the figure at this scale.

L181: “The analysis suggests. . .” this is not a complete sentence, please edit Figure 5: Please make x and y labels larger

L229: change phrase starting with “if shadows. . .” to “shadows covering dry land in the vicinity of the lake may cause an overestimation of the surface area of the lake”

L225-234: Does the cloud masking method remove cloud shadows?

L225-234: Is the influence of topographic shadowing examined? Topographic shadowing, particularly in the NY lake (in winter) could influence classification accuracy (and would not be a randomly distributed error). Even if most of the lakes specifically examined here occur in the tropics or in areas with little-to-no surrounding topography, topographic shadowing issues would likely impact the applicability of this method in other areas and therefore should be discussed

References

Avisse, Nicolas, et al. "Monitoring small reservoirs' storage with satellite remote sensing in inaccessible areas." *Hydrology and Earth System Sciences* 21.12 (2017): 6445.

Ankush Khandelwal, Anuj Karpatne, Miriam E. Marlier, Jongyoun Kim, Dennis P. Lettenmaier, Vipin Kumar, An approach for global monitoring of surface water extent variations in reservoirs using MODIS data, Remote Sensing of Environment, Volume 202, 2017, Pages 113-128.

Pekel, J., Cottam, A., Gorelick, N. et al. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422 (2016). https://doi.org/10.1038/nature20584.

Zhao, G., & Gao, H. (2018). Automatic correction of contaminated images for assessment of reservoir surface area dynamics. Geophysical Research Letters, 45, 6092– 6099.