<u>Referee #2</u>

In this study, the authors developed a hybrid remote sensing algorithm to estimate the time series of water surface areas over several lakes with rapid changes. The major novelty of this paper is that the proposed algorithm is still workable when the remote sensing images are partially covered by clouds/low-quality pixels. To enhance the capability of working under cloud condition, the authors first utilized unsupervised algorithm to classify pixels with high quality and then create a high-confidence inundation frequency (IF) image. Next, the supervised algorithm and the IF image were used to interpret masked pixels. The validation results against in situ observations indicate that the algorithm is able to monitor water surface changes with a high accuracy. This paper is well-written and easy to follow. However, there are several major concerns related to the following aspects: (1) incomplete literature review (2) the applicability of the algorithm and (3) the design of the validation.

Response
We thank the reviewer for the positive comments and hope that our responses (in **blue**) to their comments below, and the associated modifications to the manuscript (in **red**) have addressed their three specific concerns.

(1)Some efforts have been made to monitor surface water on global scale using remote sensing. It is not only necessary to acknowledge those work in this paper, but also meaningful to highlight the difference. For example, Pekel et al. (2016) developed a global water map by applying Machine Learning algorithm to Landsat images. Khandelwal et al. (2017) developed a method to monitor global lakes using MODIS images. Very similar to this paper, Zhao and Gao (2018) used an automatic approach to correct the contaminated images for assessment of reservoir surface area dynamics over 6,817 global reservoirs. They used the water occurrence map derived by Pekel et al. (2016) to correct the pixels with low quality, which is very close to the idea of IF image in this paper.

Response
We thank the reviewer for their suggestions and have rewritten the introduction to address it (also in response to comment 3 from reviewer 1). In the new introduction we provide an extensive review of the literature (now specifically referring to the mentioned studies, among others), and clarify our contribution as specifically addressing the problem of retrieving classification information from masked pixels. We agree with the reviewer that the approach from Zhao and Gao (Z&G), which we were not aware of, addresses the same problem with a comparable method, although with important differences. Z&G is indeed similar to the proposed approach in that it leverages historical inundation frequency information to infer the flooding status of masked pixels. Both approaches rely on the assumption that all masked pixels with inundation frequency above a certain threshold should be designated as flooded. However, the two approaches differ in the way they define this threshold, which varies from image to image depending on the current level of the lake. Z&G use a heuristic approach based on the shape of the histogram of unmasked flooded pixels to determine the threshold. In contrast, our approach

relies on machine learning (supervised classification) to leverage the relationship between the inundation frequency and current status of *all* unmasked pixels: flooded and unflooded. By relying on statistical relationships embedded within each image, rather than on arbitrary heuristic rules that are not able to vary across images, the proposed approach is more flexible and less reliant on user input. We now explicitly refer to Z&G in the introduction, along with the other approaches that we are aware of that address the same challenge (e.g., Avisse et al 2017 and Khandelwal et al 2017, which both rely on a DEM), and clarify the specific contribution of the proposed approach.

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

2)The applicability of the algorithm needs to be further demonstrated given the fact only a few lakes were tested in this study. Besides this, all the selected lakes are very large considering the 30-m spatial resolution of Landsat. And the majority of the global lakes are much smaller than 1 km2. It would be interesting to add a section discussing the performance of the algorithm over small lakes. We can find the lowest accuracy was obtained over the smallest lake in this study which raises a question - if the accuracy goes down when the size of lakes decreases.

Response
We agree with the reviewer and address lake size in a revised conclusion section focused on practical implications. More generally, the validation of our approach (now) focuses on investigating two key potential limitations: its sensitivity to (i) too little information in the input imagery and (ii) wrong information in the input classified imagery, both of which are now discussed in the discussion section. We use a combination of numerical experiments and targeted case studies to assess both limitations, as described in our response to ‘comment 2’ from reviewer 1.

We make the argument that lake size mainly affects the performance of our approach by decreasing the amount of information available to classify masked pixels. This effect is investigated in the revised manuscript by running a numerical experiment where a predetermined fraction of randomly selected pixels were masked from each input image. We find that the method’s ability to predict the class (water or land) of the artificially masked images is robust to an increasingly large fraction of masked images. We interpret this as indicative of the method’s applicability to small lakes, provided 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 nonetheless satisfied for the two ~1km2 lakes that we added to the validation sample in the revised manuscript. (Unfortunately, while most global lakes are indeed smaller than 1km2, very few have openly available in-situ data of their surface area at the frequency and coverage period considered in this study.)

In contrast, the method was very sensitive to classification errors in the input imagery (e.g., clouds misclassified as water, water misclassified as land, etc), which we interpret as indicative of its limitation in situations where circumstances (topography, climate) makes it difficult to systematically identify water on the input multispectral imagery. These effects might have confounded the effect of lake size in the original manuscript: lakes in NY state, where topography and climate made water detection more challenging, were also generally smaller than the other validation lakes in Texas. To disentangle these effects, we add three lakes in the revised manuscript: two small lakes (~1km2) in Texas where water detection is straightforward, and one larger lake in NY where water detection is more challenging. We find that the method performs well on the former and poorly on the latter. This is in line with the numerical experiment results and suggests that water detection, rather than lake size, stands as a more immediate limitation of the approach.

Text modification:

- Added Figure and results section on the numerical experiments
- Added validation lakes in Figure 3
- Added discussion section on the effect of “too little” information due to cloudiness, an 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 the potential and limitations of the method’s application to small lakes.

(3)To validate the results, the authors compared all remotely sensed water extents with in situ observations. However, one of the highlights of the proposed approach is to interpret the masked pixels. So I am wondering how the algorithm performs during the cloudy season when lots of pixels are masked.

Response

We thank the reviewer for their suggestion, and have added to the revised manuscript a figure for each validation lake representing lake extent estimation error against cloudiness. We find that errors are not significantly correlated to the extent of cloud cover if clouds are appropriately detected., This finding is in line with our interpretation that the approach is robust to having too little (small or heavily clouded lakes), information in the input imagery. However, our numerical experiment also indicates that errors are correlated to the prevalence of *wrong* information (misclassification of water or land) in the input imagery. This indicates that overly greedy cloud/shadow masking algorithms will yield better results compared with algorithms that underpredict cloud/shadow pixels. This also entails that challenges will arise when many cloud pixels are difficult to detect (e.g., waspy clouds or cloud edges).

Text modification

- 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 discussion section on the effect of “too little” information due to cloudiness, an 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 potential and limitations of the method’s application to lakes with high/frequent cloud cover.

Other specific comments:

Thank you for your comments, which we will be sure to address in the revised manuscript.

(1) Line24, miss a space in ‘consequenceof’.
(2) It would be better to increase the font size for all figures.

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.