Referee 1 comments and suggestions

Dear Referee 1,

We greatly appreciate the time and effort that you have dedicating to evaluating our work. We have found the feedback insightful and have carefully considered all the comments, which have helped us significantly improve the clarity, structure and quality of our paper.

Reviewer comments/suggestions are in *black*; our responses are in red.

Major comments

The language used to describe the methods (and at times, approaches themselves) differs quite a lot from the standardised across other studies producing and accessing ML workflows. Most significantly, this arises in what the papers refers to as "classification and segmentation experiments (Section 2.3, 2.4, 3.1, 3.2) which is what the wider literature would probably refer to as (hyper) parameter optimisation using a grid search and "using input features".

This is a fair point and was also mentioned by Referee 2. We have updated the text to replace references to "experiments" with language more in line with the suggestions.

*However, I have some issues with the actual results here. The first is relating to the validation issue shown above: testing of the multi-image-trained classifier is performed *on the images the classifier was trained on*, which is not a suitable test to make for the claim made. This error is repeated in Figure 5, where (to the best of my understanding) the single-image classifier is shown classifying the image it was trained on. In this context, I am unsurprised the single-image-classifier performed better. To properly validate the claim that a single image performs well in classifying other images, an entirely separate test dataset should be assessed (perhaps including data from another region to show how this can/can't be transferred geographically).*

To avoid validating the classifier on the same image or set of images used for training, we have identified suitable images from the Landsat archive to use instead, identifying four images (one from each of LT05, LE07, LC08, and LC09), as detailed in the tables below. Unfortunately, there is only a single LT04 image of the study area available, so the updated validation work and comparison can only be done using the other four sensors. Because the differences between LT04 and LT05 are minimal, however, we do not feel this impacts the conclusions drawn.

In response to this comment, and comments from both the Editor and Referee 2, we have also applied the trained classifier to an image covering an area in Patagonia. The results for this are shown in Table 6 below and demonstrate the applicability of the trained classifier to other regions.

Table 1. Images used to train the Classifier for the multi-sensor comparison.

Landsat satellite	Landsat ID	Path / Row	Cloud Cover (%)
LT04	LT04_L1TP_075090_19901230_20200915_02_T1	75 / 90	9
LT05	LT05_L1TP_075090_20070204_20200831_02_T1	75 / 90	2
LE07	LE07_L1TP_075090_20030201_20200916_02_T1	75 / 90	5
LC08	LC08_L1TP_075090_20220213_20220222_02_T1	75 / 90	5
LC09	LC09_L1TP_075090_20230224_20230308_02_T1	75 / 90	4.6

Table 2. Images used to validate the Classifier for the multi-sensor comparison.

Landsat satellite	Landsat ID	Path / Row	Cloud Cover (%)
LT04	LT04_L1TP_075090_19901230_20200915_02_T1	75 / 90	9
LT05	LT05_L1TP_075090_20070204_20200831_02_T1	75 / 90	10
LE07	LE07_L1TP_075090_20030201_20200916_02_T1	75 / 90	9
LC08	LC08_L1TP_075090_20170215_20200905_02_T1	75 / 90	5.9
LC09	LC09_L1TP_075090_20241227_20241228_02_T1	75 / 90	11.5

Table 3. Validation metrics for the classifier trained on the single image OBIA classifier.

Satellite ID	Overall accuracy (%)	$\hat{\kappa}$ statistic	PA Water (%)	UA Water (%)	$\hat{K}$ Water
Multi-class classification					
LC09-2024-12-27	95.5	0.94	99.0	99.8	0.99
LC08-2017-02-15	93.5	0.92	93.5	93.0	0.91
LT05-2003-12-10	97.0	0.95	97.1	97.5	0.97
LE07-2001-01-10	93.4	0.91	93.9	93.2	0.92
LT04-1990-12-30	93.3	0.91	93.3	93.3	0.91
Binary classification					
LC09-2024-12-27	98.5	0.98	99.8	97.9	0.97
LC08-2017-02-15	98.1	0.97	98.6	98.0	0.97
LT05-2003-12-10	98.5	0.98	99.7	98.3	0.98
LE07-2001-01-10	98.9	0.98	98.6	99.6	0.98
LT04-1990-12-30	98.4	0.98	98.7	98.9	0.98

Table 4. Validation metrics for the classifier trained on individual OBIA classifiers per Landsat sensor.

Satellite ID	Overall accuracy (%)	$\hat{\kappa}$ statistic	PA Water (%)	UA Water (%)	$\hat{K}$ Water
Multi-class OBIA classification					
LC09-2024-12-27	95.4	0.94	99.0	99.8	0.99
LC08-2017-02-15	94.0	0.92	98.0	98.5	0.97
LT05-2003-12-10	95.6	0.94	99.8	99.7	0.99
LE07-2001-01-10	96.4	0.95	99.7	98.6	0.98
LT04-1990-12-30	94.6	0.93	99.1	98.0	0.97
Binary OBIA classification					
LC09-2024-12-27	98.5	0.98	99.9	97.8	0.97
LC08-2017-02-15	98.3	0.97	99.8	98.6	0.97
LT05-2003-12-10	98.1	0.97	98.6	96.8	0.96
LE07-2001-01-10	98.2	0.97	99.8	97.8	0.97
LT04-1990-12-30	98.2	0.97	100	95.8	0.93

Table 5. Validation metrics for the classifier trained on multiple images (one per Landsat sensor) to a singular OBIA classifier.

Satellite ID	Overall accuracy (%)	$\hat{\kappa}$ statistic	PA Water (%)	UA Water (%)	$\hat{K}$ Water
Multi-class OBIA classification					
LC09-2024-12-27	92.2	0.89	96.5	97.5	0.95
LC08-2017-02-15	91.2	0.88	99.4	97.5	0.95
LT05-2003-12-10	88.3	0.84	87.0	94.0	0.90
LE07-2001-01-10	87.2	0.83	86.1	94.1	0.89
LT04-1990-12-30	84.9	0.8	84.4	92.2	0.9
Binary OBIA classification					
LC09-2024-12-27	98.3	0.97	99.8	98.8	0.97
LC08-2017-02-15	98.3	0.97	98.0	99.8	0.99
LT05-2003-12-10	98.7	0.98	96.7	98.8	0.98
LE07-2001-01-10	97.6	0.97	96.7	98.8	0.97
LT04-1990-12-30	96.6	0.95	95.6	95.9	0.95

Table 6. Validation metrics for the Patagonia image classified using the classifier trained on an image of New Zealand.

Satellite ID	Overall accuracy (%)	$\hat{\kappa}$ statistic	PA Water (%)	UA Water (%)	$\hat{K}$ Water
Multi-class OBIA classification					
LC08-2021-03-03	92.8	0.91	98.9	97	0.96
Binary OBIA classification					
LC09-2023-03-03	98.8	0.97	97.9	100	1

Additionally, it is surprising, given the ease at which this can be done in GEE, that the classifier isn't subsequently applied to a full-time stack of imagery, and data presented of example lake(s) area through time.

This is a great comment, and this is the focus of a second study that focuses on the glaciological aspects of glacial lake changes in the study region. The main aim of this study was to present the approach and how we came to choose the optimal parameters.

Sharing a way of applying the trained classifier (see comment to L603 below) would elevate this paper towards having real applicability and impact.

In response to this comment, we have exported the trained classifier as an asset for use in Google Earth Engine and will also include a detailed markdown on how to use this trained classifier in GitHub.

Line comments

L11-12 This sentence (and the following) comes a bit out of nowhere - the classification method (OBIA) is not mentioned by name until line 15, so it is unclear what method the parameters/input features/training data relate to.

Good point. We have reorganised the sentence structure here so that OBIA is mentioned first and then we mention the parameters/input features/ training data this relates to.

L36 - PBIA is mentioned as though it is distinct from approaches such as NDWI (L35-27), when in fact most applications of indices are applied in a PBAI fashion (as stated L41-43). Indeed, this section in general probably needs a better differentiation between comparative methods, split into e.g. (i) OBIA vs PBIA; (ii) indice/threshold-based vs ML based (e.g. RF); (iii) local vs cloud-based (GEE, AWS, etc.) methods. These three classes are not mutually exclusive, and each has positives and negatives.

Great points. We have split the introduction as followed:

- Pixel-Based and Object-Based Image Analysis
- Classification methods within PBIA and OBIA
- Computing environments
- Study rationale and objectives

L93 - most efficient method? Given reference to changing parameters, perhaps this is meant to mean the optimal parameters for the classification task?

Done. We have changed the wording of this sentence.

Paragraph beginning L105 - Arguably this paragraph could be removed or highly condensed - some unnecessary information (geologic history?) and climate information is largely irrelevant to the rest of the study - the second paragraph of this section, summarising regional lake studies, is more relevant.

Done. We have removed this paragraph from the manuscript.

Fig 1. - what is the data origin of the lake outlines?

The origin of the lake outlines in the figure is directly from the outlines produced in this study. This information has now been added to Figure 1.

L123-136 - For an EO specialist journal, could probably lose this preamble and start at the sentence "We use Landsat 4-9 Collection Tier 1…".

Done.

L142 - Is it necessarily within the methods to explain that these bands were renamed within the workflow? Perhaps just say that equivalent bands were used where band numbers differed between 4-7/8-9, and refer only to the band 'name' hereon (e.g. Green, SWIR1, etc.).

Done.

L148 - The NZ DEM is 8 m but was presumably resampled to match the resolution of the Landsat imagery for feeding into segmentation/RF processes. In this case, could it be equally easy to recommend e.g. COPDEM-30, and have this be globally valid?

Good point. We have added a comparison of the results gathered when comparing the COP30 and NZ DEM. The results of this comparison can be seen below.

Table 7. Validation metrics for the classifier trained using the NZ 8 m DEM compared to the Copernicus 30 m DEM.

DEM	OBIA	Satellite ID	Overall accuracy (%)	κˆ statistic	PA Water (%)	UA Water (%)	κˆ Water
NZ 8 m	Multi-class	LC09-2023-02-24	95.7	0.94	99.2	100	1.00
NZ 8 m	Binary	LC09-2023-02-24	98.7	0.98	100	98	0.97
Copernicus 30 m	Multi-class	LC09-2023-02-24	94.8	0.92	98.2	100	1.00
Copernicus 30 m	Binary	LC09-2023-02-24	97.6	0.96	99.2	98	0.97

L162 - out of the four parameters, three are referred to in prose and one by the actual parameter name within the GEE javascript API (neighborhoodSize). I would recommend consistency here.

Done. We have ensured that the four parameters have the correct Google Earth Engine API names.

L185-9 - "undertook a total of 48 (40+8) experiments" - 'experiments' should probably be better referred to as a gridded parameter sweep aiming to find the optimal number of training points and trees, followed by a second parameter sweep using optimised parameters from the first round which aimed to find the best segmentation parameters.

We have addressed this by updating the language used; however, in response to the comment below, we have also changed how this was implemented by varying all three components (training points, trees, and segmentation parameters) simultaneously, for a total of 160 different classification outputs.

L197 - I am slightly unclear as why the RF parameter sweep was performed before the segmentation parameter sweep. As the segmentation parameters will alter the input to the RF training, then changing the segmentation may alter the optimal RF values (as an addendum, what were the segmentation parameters fixed as when running the RF sweep?). The solution here would be to perform a parameter sweep on all three component simultaneously (either as an exhaustive grid search or through a random or bayesian search if optimisation is needed).

This is a great point, and we agree that optimising the RF parameters independently of the segmentation parameters may create bias of the optimisation of our approach. In response, we performed a simultaneous parameter sweep including the RF, changing in the number of feature class points and segmentation hyperparameters.

We have applied this solution to both the multi-class and binary OBIA workflows. This does not represent a different classification result but performs a more rigorous optimisation and validation of the classification approaches, ensuring that the reported results are based on all three components simultaneously.

L198 - With only 8 experiments, how were connectivity/neighbourhoodSize determined? Was this e.g. a 4x2 grid search?

This would now feed into the above and we are currently going through the results and changing our figures to address the comment above.

L207 - is there a better term than 'hillshade' to describe this data, given the well-established meaning within the literature?

We are not sure what is meant here – the "hillshade" referred to here is created using the .hillshade() method implemented as part of the ee.Terrain class.

L269 – is it perhaps fair to say that the RF model is robust to the number of trees (probably unsurprising, given the nature of RF).

Yes this would be fair to say.

Figure 5 - Assuming this is the 2023 LC09 image (it is not made clear in the caption) it is probably not surprising that the single-image-classifier performs better - it has been trained solely on this image, but the multi-image classifier has been trained on a larger set! A different image should be selected for both qualitative and quantitative assessment - indeed, the authors should be more careful throughout that training data is not being used in their model validation steps.

We have re-done this part of the analysis using the updated set of images shown in **Table 1**-**Table 5**.

Table 5 - Perhaps I miss this in the discussion, but it's probably worth including the context of GEE recently moving to limit 'free' usage for academic users. After all, a couple of minutes difference in runtime is largely inconsequential if one is producing a classifier that will be used thousands of times - but a double in EECU-seconds may be highly significant if one is training many models in a 'free' plan.

Thank you for pointing this out to us. We have now updated this in the manuscript.

Comment: L419 - why do higher sample sizes lead to a reduction in classifier accuracy? Is this simply overfitting or something else?

This is likely to do with overfitting.

Fig 7 and associated discussion - Is a solution to train separate RF models for the older and newer Landsats?

This is likely a good solution as it would match the radiometric resolution in the training and testing of our OBIA approach.

L603 - Nice presentation of figure and data code within the GitHub. However, I'm less clear on how I can access and use the actual model trained here, although perhaps I am just missing something. If the classifier is indeed broadly applicable as stated, it would be useful to have an additional GEE code (and perhaps explanatory tutorial included in the Markdown documentation) showing how users can apply the trained model to a new region.

We agree and will ensure that we export the trained classifier in Google Earth Engine and provide an in-depth explanation on how you would be able to apply this yourself to a new region.