Responses to AC1

Dear referee,

We would like to express our sincere appreciation for your careful reading and helpful comments to improve the quality of this paper. Such comments are extremely valuable and helpful for revising and improving our paper. We apologize for the problems in our manuscript, and we have fixed them in the revised version. Our point-by-point responses to your comments are given as follows. The sentences in italic and underline are the comments.

Main comments

✧ Comment 1:

Page 2 lines 39-44 - The discussion about semi-automatic methods is unclear. Please point out what makes these methods semi-automatic. For example, are these methods that use specified thresholds or feature as part of their workflow? Why are they more regionally restrictive than automatic methods. Please revise this part.

Respond:

Thank you very much for your comment and valuable suggestion. The intraclass heterogeneity of glacial lakes is large, and glacial lakes in different regions tend to show greater differences. Therefore, when using traditional semi-automatic methods, it is necessary to manually adjust the index threshold or select more appropriate features according to the characteristics of glacial lakes and ground backgrounds in different regions. Besides, traditional machine learning relies on mathematical models for classification, and artificially set glacial lake labels and parameters have a great impact on the final results. For automatic methods such as deep learning, the computer automatically extracts features of the glacial lake and automatically updates parameters without human intervention. The following is the revised part.

Re-edit:

Due to the different elements contained in the water body of the glacial lake and the depth of the glacial lake, the glacial lake has a large intra-class heterogeneity, and different spectral information is displayed on the optical remote sensing image. Although these semi-automatic extraction methods are widely used, the operation is manually dependent and regionally restrictive, limiting the promotion and application on the global/hemispheric scale. For example, the glacial lakes in the Mount Everest area are mostly blue or green, while the glacial lakes developed in the deep valleys of the eastern section of Nyainqentanglha are closer to black. Therefore, when using traditional machine learning methods, it is necessary to artificially select appropriate features according to the characteristics of the glacial lake and ground background in the area.

✧ **Comment 2:**

Page 2 line 56 - Why is the method in Zhao et al. 2018 only applicable to Landsat images (is this due to the bands available, if so, please state this as it is more specific).

Respond:

The method uses the SWIR band of Landsat images, while most high-resolution images have only four bands of red, green, blue and near-infrared, and Google Earth images have only three bands. Therefore, it is written “only applicable to Landsat images”. But after your kind and valuable reminder, we realized that the ten-meter-level Sentinel images also have the SWIR band, and it may be more appropriate to change it to "only applicable to remote sensing images with multiple bands (especially SWIR) such as Landsat images". The following is the revised part.

Re-edit:

It is only applicable to remote sensing images with multiple bands (especially SWIR) such as Landsat images, while most high-resolution images have only four bands of red, green, blue and near-infrared.

✧ **Comment 3:**

Page 2 line 67 – ‘with NDWI as the spatial attention’ - what does it mean to have a feature as the attention mechanism? Or do you mean something different here? I could not find this in the paper (He et al. 2021). Did you mean to reference the paper 'J. Wang, F. Chen, M. Zhang and B. Yu, "NAU-Net: A New Deep Learning Framework in Glacial Lake Detection," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 2000905, doi: 10.1109/LGRS.2022.3165045'?

Respond:

For “with NDWI as the spatial attention”, we mean to reference the paper "NAU-Net: A New Deep Learning Framework in Glacial Lake Detection". We are very sorry to cause you trouble here, we have modified the language expression here, and hope that it can explain clearly after the modification.

Re-edit:

He et al. (2021) added a space attention mechanism into the skip connection of U-Net to focus on glacial lakes. Wang et al. (2022b) proposed NAU-Net with NDWI as the spatial attention, which guided the network to pay more attention to the glacial lake information of low-level features and solved the problem of the area difference between the area occupied by positive and negative samples.

✧ **Comment 4:**

lines 124 -125 I usually use data, like Landsat, downloaded directly from e.g. NOAA or similar sources, and am not very familiar with Google Earth Engine. Other readers of this journal may be similar. I don't follow specifically what images are used or what the different levels (e.g. levels 14 to 19 of Google Earth images) mean - could you please elaborate?

Respond:

Thanks for your kind suggestion. Google Earth images is not a single data source, but a data collection of various aerospace imagery (Landsat, QuickBird, etc.) and aerial photography. Different levels of Google images correspond to different spatial resolutions, the higher the level, the higher the spatial resolution. For example, the spatial resolution of Google Earth imagery at level 19 is 0.14 meters, and at level 14 it is 4.45 meters. The following is a supplementary introduction to Google Earth imagery.

Re-edit:

Google Earth imagery is a data collection of various aerospace and aerial remote sensing images, including Landsat , QuickBird, IKONOS images and other data. Different levels of Google Earth images have different spatial resolutions, and the higher the level, the higher the spatial resolution.

✧ **Comment 5:**

line 134 Could these image tiles randomly selected for validation be part of the same scene as those used for training data, and if so, how do you think this may impact your scores?

Respond:

Thanks for your kind comment. In the glacial lake dataset, all image tiles from different continents are mixed together. The glacial lake samples were collected from glacial lakes in different scenes to ensure the richness and comprehensiveness of the glacial lake samples. Therefore, the training samples (80%) and validation samples (20%) randomly selected in the dataset during model training do not specifically correspond to a certain scene, but the overall training and validation results. We are very sorry to cause you trouble here, we have modified the language expression here, and hope that it can explain clearly after the modification.

Re-edit:

Finally, a total of 15376 samples with a size of 256×256 were obtained, out of which 20% of image tiles were selected as validation data randomly (Table 1).

✧ **Comment 6:**

lines 147-148 - For this audience I think more information is needed as to what SLIC and Dense CRF are, and why they are chosen.

Respond:

Thank you very much for your comment and valuable suggestion. The following is the revised part.

Re-edit:

In the pixel-based semantic segmentation, the outline of the glacial lake is not refined enough, which does not fit the actual smooth edge of the glacial lake. The Simple Linear Iterative Clustering (SLIC) algorithm could fuse the rough result of semantic segmentation with the edge information of superpixel

segmentation to enhance the integrity of the glacial lake and improve the edge segmentation. The Dense Conditional Random Field (DenseCRF) uses the constraint relationship between pixels to encourage similar pixels to be assigned the same label, while pixels with large differences are assigned different labels to obtain accurate glacial lake outlines. In this paper, after semantic segmentation, two-level optimization combined SLIC and DenseCRF was used to achieve refined extraction of glacial lake outlines (output 2 and output 3 in Fig. 3). By the way, these two optimization methods can also be used separately to implement single-level optimization.

✧ **Comment 7:**

In Figure 3 - Are the arrows that indicate `First optimization ' and `Second optimization' steps (e.g., in a code) or is the first optimization the steps from labels/GEE images to output 1, and if so would it be better to put a box around this part and call it `first optimization'. Similar comment applies to the 'second optimization'.

Respond:

The arrows in Figure 3 indicate "First optimization" and "Second optimization" steps. Based on your valuable suggestions, we added boxes to make the structure diagram more clear. Thanks again for your kind comment.

Re-edit:

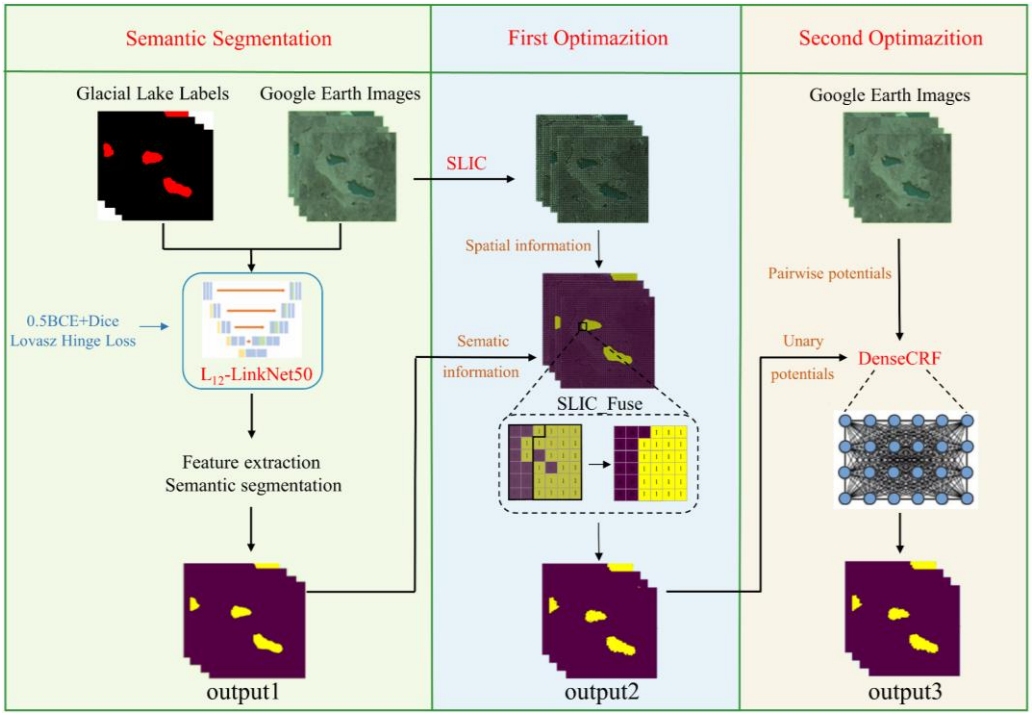

Figure 3. Structure diagram of glacial lake extraction strategy in this study. All copyrights Image ©Google Earth 2020.

✧ **Comment 8:**

lines 162-165 The reader might wonder what is so special about using LinkNet. It's a little hard to follow the motivation here - LinkNet is able to extract deep/complex features with fewer weights than a standard U-Net - is this essential for this problem? Or does it have to do with the addition property of LinkNet when the features from the encoder and decoder are combined?

Respond:

Thank you for your suggestion, we have explained in the text corresponding position. After multiple down-sampling by Encoder, some spatial information is lost, and it is difficult to recover these lost spatial information in the Decoder part, which is not conducive to the extraction of glacial lakes. The commonly used U-Net network uses concat operation to combine the feature map obtained by the encoder part with the feature map corresponding to the decoder. However, the size of the feature map corresponding to the two layers is inconsistent, so the feature map obtained by the encoder needs to be cut to complete the splicing, and some effective information is lost. In the LinkNet network, the size of each layer feature map corresponding to Encoder and Decoder is the same, and the addition method is used to combine the features, and the shallow features are re-learned without increasing the parameters, so as to effectively restore the spatial information of the glacial lake.

Re-edit:

The LinkNet network (Chaurasia and Culurciello, 2017) uses ResNet18 (He et al. 2016) as the backbone of the U-Net (Sathananthavathi and Indumathi, 2021). In the LinkNet network, the size of each layer feature map corresponding to Encoder and Decoder is the same, and the addition method is used to combine the features, and the shallow features are re-learned without increasing the parameters, so that the spatial information of the glacial lake can be effectively restored, which has a lightweight structure and fast calculation speed.

✧ **Comment 9:**

Figure 4 - tell the reader what the difference is between a deconv is in comparison to a transposed conv and why you have this distinction in your network.

Respond:

Thanks for your kind comment. There is no difference in method between transposed convolution and deconvolution. In the pytorch code, both are nn.convtranspose2d(), only the parameters and uses are different. The following are revised notes to Figure 4.

Re-edit:

Note: In the Decoder, Conv(1×1) is responsible for reducing the number of channels(×1/4), and Deconv(3×3) only changes the size of the feature map (×2). After the decoder, Transposedconv (Deconv(4×4)) will reduce the number of channels (×1/2) and expand the size of the feature map(×2).

✧ **Comment 10:**

Both sections 3.2.1 and 3.2.2 don't give the reader an idea as to why a superpixel segmentation algorithm is needed, or why a conditional random field are chosen. This information is also not clear earlier (in introduction or background material). Hence it's hard to read these two sections and extract useful information. If these details aren't discussed further, the overview (i.e., why these two methods are chosen) could be put earlier in the manuscript, while the details in 3.2.1 and 3.2.2 could be put in an appendix.

Respond:

Thank you very much for your comment and valuable suggestion. The reason for choosing a conditional random field has been updated in the revised part.

Re-edit:

In the pixel-based semantic segmentation, the outline of the glacial lake is not refined enough, which does not fit the actual smooth edge of the glacial lake. The Simple Linear Iterative Clustering (SLIC) algorithm could fuse the rough result of semantic segmentation with the edge information of superpixel segmentation to enhance the integrity of the glacial lake and improve the edge segmentation. The Dense Conditional Random Field (DenseCRF) uses the constraint relationship between pixels to encourage similar pixels to be assigned the same label, while pixels with large differences are assigned different labels to obtain accurate glacial lake outlines. In this paper, after semantic segmentation, two-level optimization combined SLIC and DenseCRF was used to achieve refined extraction of glacial lake outlines (output 2 and output 3 in Fig. 3). By the way, these two optimization methods can also be used separately to implement single-level optimization.

✧ **Comment 11:**

lines 230 'full-text' information - can you place this in the context of the present study?

Respond:

The ' full-text ' here means all the pixels in the image. The category of each pixel in the DenceCRF algorithm is related to the category of all other pixels in the image. DenceCRF connects all pairs of individual pixels in the image, enabling greatly refined segmentation and labeling. It means the same as the global context information in the following text.

✧ **Comment 12:**

No details about RF or SVM are given. We need more information to understand if this is a reasonable comparison. It these two methods were used in a previous study by the authors then they could refer to this here. These two could also be omitted since there is also a comparison to UNet and EfficientNet UNet if adding that information would make the manuscript too long and the emphasis is not on feature learning vs feature engineering.

Respond：

Thank you for your advice. We first use SLIC for superpixel segmentation, and perform feature extraction of pixel blocks to obtain training samples, and then use SVM and RF two traditional machine learning algorithms for classification. For the parameter settings of these two algorithms, we supplement them in the revised part. In the SVM classifier, the penalty coefficient (C) was set to be 12, the kernel to be radial basis function (rbf), and the gamma in kernel function to be 0.187. In the RF classifier, the number of decision trees was set to 150, and the number of features is set to 2.

Re-edit:

In the SVM classifier, the penalty coefficient (C) was set to be 12, the kernel to be radial basis function (rbf), and the gamma in kernel function to be 0.187. In the RF classifier, the number of decision trees was set to 150, and the number of features is set to 2.

✧ **Comment 13:**

lines 271-273 - Show in your figures where the false detections and shadows are. For example, figure5 does not say what the red and yellow circles are referring to. A similar comment applies to the multi-color circles in Figure 6.

Respond：

Thank you very much for your comment and valuable suggestion. The circle represents the difference between the results extracted by different methods and the real label (yellow : real label, red : extraction result), guide readers to focus here. In order to improve readability, we have unified the colors.

Re-edit:

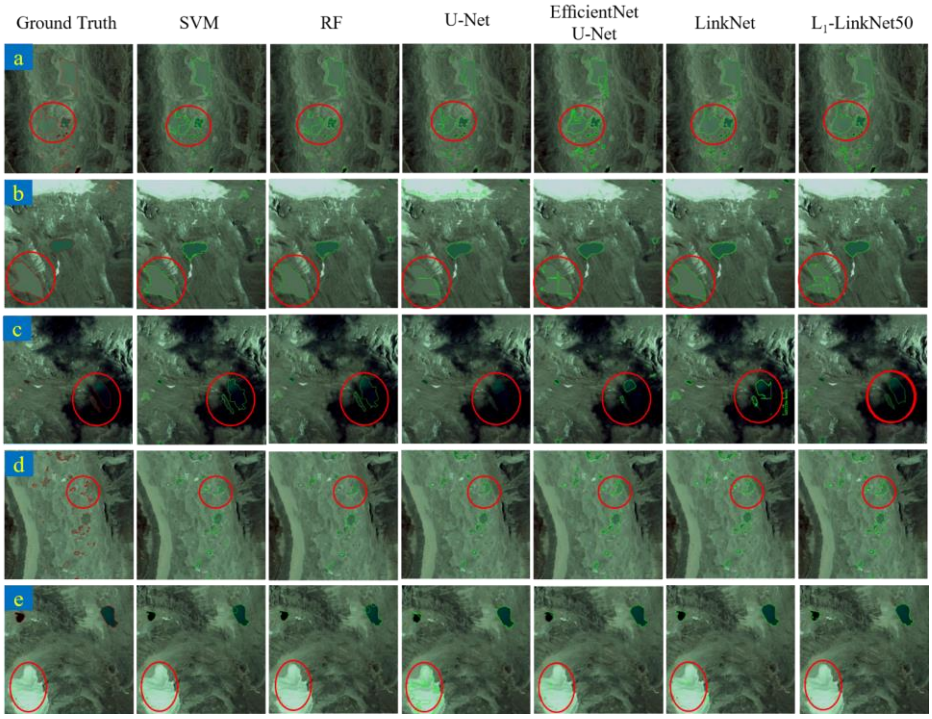

Figure 5. Performance comparison of different models for the glacial lake extraction. All copyrights Image ©Google Earth 2020.

Note: Five regions of the same size (2048×2048) were chosen based on Google Earth images (0.52 m). The red vectors are the boundary of the truth glacial lakes, and the green vectors are the boundary of the extraction result, including areas with glacial lakes of complex outlines (a), the inconsistent color of water bodies (b), mountain shadows (c), and areas with multiple small glacial lakes (d), ice and snow (e).

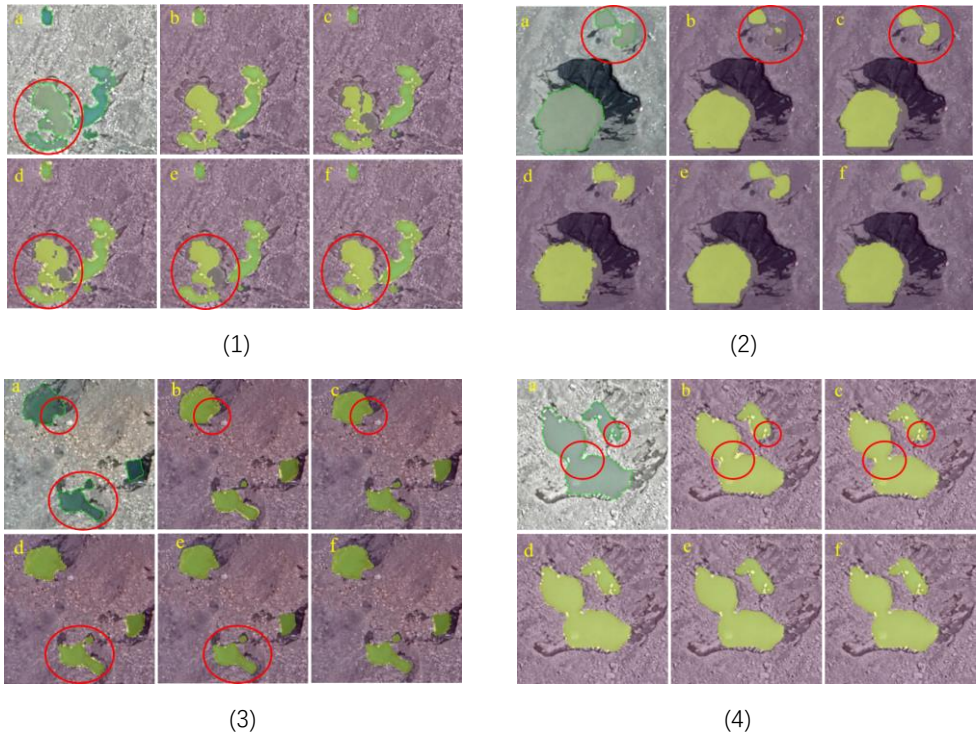

Figure 6. Comparison of glacial lake identification results based on Google Earth images (0.52 m) under different optimization conditions. All copyrights Image ©Google Earth 2020.

Note: Ground Truth (a), L_1-LinkNet50 (b), L_{12}-LinkNet50 (c), L_{12}-LinkNet50-SLIC (d), L_{12}-LinkNet50-DenseCRF (e), and L_{12}-LinkNet50-SLIC-DenseCRF (f). The green vectors are the boundary of the truth glacial lakes, and the yellow masks are the semantic segmentation result.

✧ **Comment 14:**

I am not sure if contribution (2) on page 3 is considered a contribution since this loss function has been used in other studies. However, it may be the first time this has been used for segmentation of small objects? For example, a related (but different) approach was taken for the problem of sea ice floe segmentation ('Nagi, A.S.; Kumar, D.; Sola, D.; Scott, K.A. RUF: E_ective Sea Ice Floe Segmentation Using End-to-End RES-UNET-CRF with Dual Loss. Remote Sens. 2021, 13, 2460. https://doi.org/10.3390/rs13132460') but without the Lovaz loss, which seems to make a significant different here. It might be worth pointing this out (perhaps in the conclusions) since the small lakes are somewhat similar to ice floes in that they are irregular small objects in background state.

Respond:

Thank you again for your positive comments and valuable suggestions. After using the loss function of 0.5BCE + Dice to train the network, we use Lovasz Hinge Loss to fine-tune the LinkNet50 network. Our difference lies in the use of a two-step constrained loss function and training strategy. Therefore, we modified the expression in the manuscript.

Re-edit:

(2) To alleviate the negative impact of unbalanced positive and negative samples on the network extraction for glacial lake features, a two-step constrained loss function and training strategy were proposed with Resnet50 as the backbone.

✧ **Comment 15:**

the conclusions refer to the high F1 score in 'the study area' but not (more generally) the results in the QNNR - why is this?

Respond:

Thank you for your suggestion, we are not comprehensive summary of the article. We have made modifications to the corresponding part.

Re-edit:

The F1 Score in the study area reaches more than 90%. At the same time, it is applied to a wider range of QNNR. Due to the misjudgment of the small glacial lake in the shadow, the F1 score is reduced, but it also reaches 82.49 %.

Minor comments

✧ **Comment 1:**

In the abstract the authors refer to `Google Earth' images - please be more specific about what images are used. In addition two different image resolutions (2.11 m and 0.52 m) are referred to but it's not sure why two different resolutions are used. They also refer to `the study region' and then later (in the abstract) `Qumolongma National Nature reserve', which adds to the confusion. reply:

Respond：

Thanks for your kind suggestion. Google Earth images is not a single data source, but a data collection of various aerospace imagery (Landsat, QuickBird, etc.) and aerial photography. Different levels of Google images correspond to different spatial resolutions, the higher the level, the higher the spatial resolution. 2.11 m and 0.52 m are the highest spatial resolutions of Google Earth images available in this area. The study region is used to verify the feasibility of the method. After proving that the method is feasible, the glacial lake is extracted from a larger area (Qumolongma National Nature Reserve), which belongs to the application area.

✧ **Comment 2:**

abstract line 20 - if pixel spacing is 2.11 m then wouldn't 6×6 pixels be smaller than 160m^2.

Respond：

Thank you for your careful reading and apologize for our inappropriate expression. We have made changes in the corresponding position.

Re-edit:

The area of the minimum glacial lake that can be extracted is 160 m2 (less than 6×6 pixels).

✧ **Comment 3:**

page 2 line 1 'time and vigor' - A different word could be used than vigor - resources?

Respond：

We sincerely thank the reviewer for careful reading. As suggested by the reviewer, we have corrected the 'time and vigor' to 'time and resources'.

Re-edit:

Still, it costs lots of time and resources, which is challenging to meet the needs for large-scale glacial lake identification.

✧ **Comment 4:**

page 2 line 62 'EfficintNet' typo

Respond:

We are very sorry for our careless mistake and it was rectified.
Re-edit:
Qayyum et al. (2020) used the pre-trained EfficientNet as the backbone of the U-Net to map glacial lakes,

✧ **Comment 5:**

'better result' - better than what?

Respond:

We are very sorry for our unclear expression. and it was rectified.
Re-edit:
Qayyum et al. (2020) used the pre-trained EfficientNet as the backbone of the U-Net to map glacial lakes, which achieved a better result than the original U-Net, RF and SVM classifiers in high-resolution glacial lakes extraction.

✧ **Comment 6**

page 2 line 69 `area difference between positive and negative samples' - do you mean the difference between the area occupied by positive and negative samples (with there being more area for one than the other)?

Respond:

Thank you for your careful reading and suggestions, and sorry that our expression is not clear, we have made changes.
Re-edit:
Wang et al. (2022b) proposed NAU-Net with NDWI as the spatial attention, which guided the network to pay more attention to the glacial lake information of low-level features and solved the problem of the area difference between the area occupied by positive and negative samples.

✧ **Comment 7**

line 98 - remove 'Besides' before 'no large rivers', for example change to `There are no large rivers in the study area'.

Respond：

We think this is an excellent suggestion and we have made changes to this.
Re-edit:
There is no large rivers in the study area.

✧ **Comment 8**

It would be helpful to show the QNNR area in Figure 1 and then add text to the figure caption indicating which area is used for train/test and how the evaluation for the QNNR is done different that the entire study region (if I follow correctly only inference was done for the QNNR).

Respond：

Thank you for your advice. The data set used for training and testing in the paper contains samples from multiple regions in Asia, South America, North America and Europe. After the model training is completed, the glacial lake is predicted in the area of Figure 1. The analysis shows that the L12-LinkNet50-SLIC-DenseCRF model has the best effect, and then the model is applied to QNNR to test whether it can be applied to a large area. If the entire dataset location is displayed, a global-scale base map needs to be drawn, so that our study area is difficult to be found in the map. At the same time, since the QNNR boundary coincides with most of the county-level administrative boundaries, we did not add the QNNR vector to the figure. We annotated the same county name in Figure 7 (main picture) and Figure 1 (lower left), and the two images can be connected through the county.

✧ **Comment 9**

The inset in Figure 7 is far too small.

Respond：

Thank you very much for your correction and patience. We have noticed this problem and modified it at the corresponding location.
Re-edit:

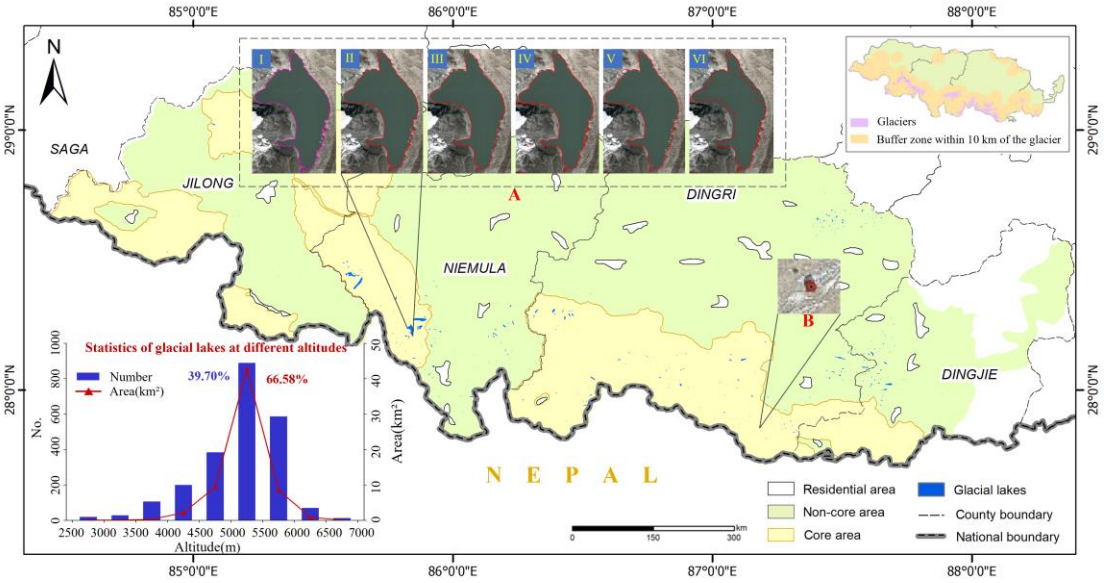

Figure 7. Glacial lake extraction result in the QNNR based on Google Earth images in 2020 (2.11 m). All copyrights Image ©Google Earth 2020.

Note: A and B are the largest and smallest glacial lakes extracted in this region, respectively. The purple vectors are the boundary of the truth glacial lakes, and the red vectors are the boundary of the extraction result. Ground Truth (I), L_{12}-LinkNet50-SLIC-DenseCRF 2.11m (II), Inventory data of glacial lake 30m (III), ESA World Cover 10m (IV), Esri Land Cover 10m (V), FROM-GLC10 10m (VI).

✧ Comment 10

line 366 - I am not sure the reader will know what the 'single-variable' method is.

Respond:

Thank you for your advice. What we want to express is that the two parameters superpixel blocks and the compactness that SLIC needs to adjust are tested separately. We have supplemented the manuscript. This means that when the compactness remains constant but the number of superpixel blocks varies, the optimal number of superpixel segmentations is selected. Similarly, the setting of compactness follows the same approach. We replace it with the control variable method.

Re-edit:

were obtained through multiple experiments by the control variable method based on sub-meter-level images.