RC1

The proposed manuscript “Global Attention of Transformer Empowers Montane Periglacial Lake Identification” seeks to advance the remote-sensing based detection and classification of montane periglacial lakes. Accurate inventories of these lakes are of particular relevance as these lakes are indicators of climate change, important sources of fresh water, and pose geohazard risks through GLOFs. The authors identify three main challenges in the remote-sensing based detection of periglacial lakes, which are the difficult detection of very small lakes, spectral confusion due to topographical shadows and similar land surface classes, and the discrimination between glacial and non-glacial lakes. To address these challenges, the authors propose a two-stage classification approach, in both of which Vision Transformer (ViT) models replace more established models.

First, lakes in a Himalayan study region are detected from a Sentinel-2 mosaic using image segmentation. For this the authors propose the ViT model Mask2Former. Second, the identified lake shapes are analysed in their original environmental context to semantically classify them as either glacial or non-glacial. For this task, the authors propose the Swin Transformer v2 model. The models are trained in one region and applied and tested in a second to avoid overfitting and ensure transferability. The model results are compared to those of different established convolutional neural networks (CNN) architectures, and the proposed framework appears to yield better results throughout. The final mapping product for the validation region is furthermore compared to two different lake mapping products. The new mapping approach detects a significantly larger amount of lakes thank the comparison datasets, which the authors attribute to the ability of their framework to detect particularly small lakes.

General comments

The manuscript has a clear approach, is generally well structured and concise. The methodology of comparing a newly developed framework to existing ones is suitable. The discussion does well in explaining the performance of ViT compared to CNN based on the different model architectures. The presented results are in so far relevant, as they seem to be a significant improvement in comparison to established lake mapping methods (e.g. the U-Net) in montane areas. Even though the study is driven rather by a methodological instead of a geoscientific research question, I feel that with some rework it can be a valuable contribution to the cryosphere research community as it demonstrates a way to generate comprehensive inventories of periglacial lakes.

However, I think the paper needs some major revisions before publication. A major point is that the authors need to elaborate more on their methodology to facilitate transparency of their experiments and reproducibility. More details and explanation would help the geoscientific community to better understand the selected model architectures and configurations. More specific feedback on these issues can be found in the Specific Comments below. Furthermore, I strongly encourage the authors to share the lake labels used for training and testing in an open repository, not only for transparency but also to bolster the credibility of their results. Otherwise, it will be impossible to verify these. The same is true for the programming code of the applied models should they have been modified from their original source. Finally, I find the part of the discussion that addresses the confusion of glacial and non-glacial lakes in close proximity to glaciated areas to be insufficient. This could be improved by a similar analysis as presented in the results section. Again, more specific feedback on that can be found in the Specific comments.

Response: We thank you for the constructive assessment and for recognizing the potential value of this work for cryosphere research. In response to your general and specific comments, the manuscript has been carefully revised to improve methodological transparency and reproducibility, including clearer descriptions of model architectures, training strategies, and evaluation procedures. The discussion has been strengthened to better address lake type confusion in near-glacier environments, and the training and validation lake labels have been made publicly available to support transparency and independent verification.

Detailed point-by-point responses to your specific comments are provided below.

Specific comments

Title: I feel the word “empower” is too strong, as the proposed method rather advances/improves the already working detection of periglacial lakes.

Response: We agree that the term “empower” conveys a stronger implication than intended. In the revised manuscript, it has been replaced with “enhancing” to more accurately reflect that the proposed method advances and improves an already effective periglacial lake detection framework through the incorporation of global contextual information enabled by the Transformer architecture.

In addition, considering the concerns raised by another reviewer regarding potential ambiguity in the terms “periglacial lake” and “global attention,” the title has been comprehensively revised. The final title of the manuscript is:

“Enhancing Lake Identification in Periglacial Environments by Leveraging the Global Context of Transformers.”

L52-53: What is the reasoning for this exact lake size threshold? Is it sensor resolution? Is it the low relevance of lakes of such small a size? Or from a different perspective: Why is it important to also include these small lakes and develop a method, which is able to detect these? I think it is worthwhile to address this, as the proposed methods later shows its strengths at exactly this lake size.

Response: The area threshold of 0.001 km^2 adopted in many previous studies (e.g., Nie et al., 2017; Chen et al., 2021) is primarily constrained by sensor spatial resolution (e.g., 30 m Landsat imagery) and practical considerations for large-scale mapping, where smaller water bodies are more easily confused with shadows, snow cover, or ephemeral surface water in complex alpine terrain. As a result, inventories based on these datasets have typically excluded lakes below this threshold.

In the revised manuscript, this rationale has been explicitly clarified in the Introduction (Lines 60–63).

L68: What do you mean by "adaptive feature selection" in a Machine Learning context?

Response: By "adaptive feature selection," we originally referred to the ability of machine learning methods to learn and weight informative features during training, thereby enhancing robustness to environmental variability. To avoid ambiguity and potential misunderstanding in a machine learning context, this wording has been clarified in the revised manuscript (Lines 78-79).

L101: What is "Hydroformer"? Is it a CNN? How does its architecture compare to the other introduced methods. Briefly elaborate.

L101-104: To be consistent with the sources you cited before: Could you briefly add in which spatial context (location, scale) the two studies cited here were conducted.

Response: To clarify the nature and scope of the cited methods, we have revised this paragraph to explicitly distinguish their model architectures and application contexts. Hydroformer (Hou et al., 2024) is now clearly identified as a Transformer-based temporal sequence model designed for lake level reconstruction, rather than an image-based CNN or vision Transformer, and its global evaluation scope across 50 lakes of varying sizes is explicitly stated. In addition, the spatial context of LEFormer (Chen et al., 2024) has been clarified by specifying its evaluation on global surface water datasets and lakes in the Qinghai–Tibet Plateau.

These clarifications have been incorporated directly into the revised manuscript to improve consistency, readability, and comparability among the cited studies (Lines 111–116).

L107: I think what's missing here is an overview about which specific shortcomings of the ViT studies cited before the proposed approach in this paper is supposed to address. Is it just the lack of application of ViTs for the detection of periglacial lakes? I can see, that ViTs have been applied before to detect lakes (and other surface features) in different contexts, but what in the cited studies makes the authors claim that ViTs are particularly suitable for this type of setting (montane periglacial)? I very much agree that it is worthwhile to investigate the suitability of ViTs for the proposed task, but the introduction chapter could be improved by providing some stronger arguments why particularly ViTs are promising.

Response: We agree that the Introduction benefits from a clearer articulation of how the cited Vision Transformer (ViT) studies motivate the proposed approach and why ViTs are particularly promising for montane periglacial lake identification. In the revised manuscript, we have strengthened this argument by explicitly synthesizing the limitations of prior ViT applications and clarifying the remaining research gap.

Specifically, we now emphasize that while existing studies demonstrate ViT's capability to integrate multimodal data and model complex spatial contexts, their application has not yet been extended to the detection and classification of size-heterogeneous lakes in alpine periglacial environments characterized by strong terrain-induced spectral and spatial complexity. This clarification has been incorporated directly into the Introduction (Lines 116–119).

L110: The claim that the study "elucidates the underlying physical mechanisms" (of what?) is too bold. This is not at all addressed in the study.

Response: We agree that the claim of "elucidating the underlying physical mechanisms" was overstated and not supported by the scope of this study, which focuses on methodological evaluation rather than direct investigation of physical processes. In the revised manuscript, this statement has been corrected accordingly.

Specifically, the text now emphasizes that the proposed framework is designed to systematically evaluate the feature representation advantages of Vision Transformer models over CNNs in complex environments, and that the results provide methodological insights for the precise identification and classification of diverse lake types, thereby supporting more consistent mapping and characterization of periglacial landforms and related cryospheric features (Lines 121–125).

Figure 1: The figure indicates an accuracy assessment on the test data of the deep learning dataset. However, there is no arrow connecting back from the accuracy assessment to the two models. Were these models tuned and optimized or just used “out-of-the-box”? This should be also addressed in the text.

Response: The workflow in Figure 1 is intentionally linear and reflects the actual modeling procedure. Both models were trained using a unified training strategy, and the held-out test set was used exclusively for final accuracy assessment. No iterative optimization, retraining, or model adjustment based on test results was performed, which is why no feedback arrow from the accuracy assessment to the models is shown in the workflow.

Based on this modeling design, the Methods section has been updated to explicitly state that no iterative optimization or feedback based on test performance was conducted (Lines 250–251).

Figure 2: The third panel of the map (the overview) would be much more insightful if it provided a shaded relief of the topography. This way, readers not familiar with the region would be better able to understand the setting of the two study regions within the larger topographical context. Consider also to zoom-in a little bit (not too far) to the Himalayas and surrounding mountain ranges themselves. Too much space in this panel is wasted on regions which are not important to this study (Siberia, Australia, Indonesia etc.)

Response: We agree with this suggestion. In the revised manuscript, the overview panel in Figure 2 has been updated by adding shaded relief to depict the topographic context more clearly. The map extent has also been adjusted to zoom in on the Himalayas and surrounding mountain ranges, reducing emphasis on regions not directly relevant to this study.

L166: You only use imagery from a single season. As a training dataset should be diverse to reflect a wide range of environmental conditions you should provide a good explanation why you focus on this limited time frame.

Response: We agree that the use of single-season imagery requires explicit justification. In the revised manuscript, this point has been clarified in the Discussion by explaining that the analysis is based on Sentinel-2 imagery acquired during the ablation and summer–early autumn period, which is optimal for lake visibility and annotation reliability, while also acknowledging that this choice limits the model's ability to explicitly account for seasonal variability.

This limitation and its implications are now explicitly stated in Section 4.3 (Lines 539–541).

L166: During compositing, how do you account for intra-annual variability of the environment and particularly lake areas? You say you favor snow-free conditions with maximum lake extent (which is totally reasonable) but how do you control that this is reflected in the composite?

Response: Intra-annual variability in lake extent is an important consideration when generating seasonal composites, particularly in alpine periglacial environments where short-term snow cover and surface conditions can vary substantially. For this reason, we restricted the analysis to the ablation and summer–early autumn period and adopted a median compositing strategy, which favors persistent surface water signals while suppressing transient features such as clouds and short-lived snow cover. At the same time, we recognize that this approach does not strictly guarantee the absolute maximum extent for every individual lake.

Based on this consideration, the revised manuscript now provides a clearer explanation of how intra-annual variability is handled during compositing and explicitly states the associated limitation in the data description section (Lines 182–191).

L178: What is the point of upsampling 10m/30m resolution input data to 5 m? Without any additional very high-resolution data there is no information gain. Why not just stick with 10m? In fact, because the input imagery into the ViTs is tiled into tiles with a fixed number of pixels (256x256), you might be losing a lot of spatial context with the higher resolution, don't you?

Response: We acknowledge that upsampling the input data to 5 m resolution does not introduce additional spatial information. This step was therefore not intended to increase information content, but to regularize spatial sampling for downstream segmentation. Finer sampling facilitates smoother boundary representation for ultra-small lakes, which are highly pixelated at the native 10 m resolution, thereby improving boundary delineation during segmentation. With respect to spatial context, the resulting 256×256 tiles at 5 m resolution still cover approximately 1.64 km^2, which provides sufficient local context given that the vast majority of lakes in the study area are much smaller in extent.

Based on this consideration, the methodology section has been revised to explicitly clarify the rationale for resampling, the absence of information gain, and the retained spatial context of the tiled inputs (Lines 201–206).

L180ff: Training labels: Generating training labels is always a crucial process in ML/DL approaches. If two different experts were responsible for creating these labels, could you elaborate on any measures taken to ensure consistency between the labels? Also, I feel it would be a huge benefit to the community to make the training and validation labels available to the open public.

Response: Training label generation and consistency control are critical steps in this study. Labels were generated through detailed visual interpretation using RGB imagery supplemented by NDWI, following the glacial lake classification system of Yao et al. (2018). The interpretation was conducted by one researcher experienced in glacial lake studies and independently checked by a second experienced researcher. Potential discrepancies in lake boundaries and lake-type assignments were identified through side-by-side inspection and resolved by consensus with reference to high-resolution imagery. This procedure was designed to ensure consistency and reliability across the training and validation labels (Lines 207–214).

In addition, the training and validation datasets, including manually interpreted lake outlines and lake-type labels, have been made publicly available through the National Tibetan Plateau Data Center, with a permanent DOI provided in the Data availability section.

L184: How were the data standardized? Which method did you use?

Response: Input data standardization was performed using Z-score normalization, in which each band is scaled to have zero mean and unit variance prior to model training. This choice was made to ensure numerical stability and consistent feature scaling across input variables. This has now been explicitly stated in the manuscript (Lines 214–215).

L195ff: As this part is very technical ML/DL language, I would recommend some reworks to cater to the geoscientific community of this journal. Specifically, I'd like to see some elaboration on how the different components/features (e.g. multi-feature extraction, self/cross attention) of the two architectures are beneficial to the tasks of segmentation and classification of periglacial lakes in a montane setting. For example, which of the challenges described in the introduction section are addressed by choosing these model architectures and configurations.

Response: We appreciate this helpful suggestion regarding the presentation of the model architectures. The architectural descriptions have been revised to better align with the

geoscientific audience of the journal. The revised Methods section now explicitly relates key components of the two architectures to the specific challenges of periglacial lake segmentation and classification in complex montane environments, as outlined in the Introduction.

These revisions have been incorporated into the Methods section (Lines 226–246).

Methods-Section: The methods section misses an entire sub-section on the additional models used for model comparison, i.e. U-Net, DeepLab V3+, ResNet, and EfficientNet. Although this section does not need to be as detailed as the (revised) section 2.4, some basic information is indeed required, such as reasoning for the choice of the comparative models, proper citation of the sources of the models, configuration of the input data for these models, and essential model hyperparameters. The reader must be able to reproduce the experiments the authors performed.

Response: The Methods section has been revised to explicitly document the additional models used for comparison. The revised text now describes the rationale for selecting the comparative CNN-based models, provides appropriate citations, and specifies the unified input configuration and essential training hyperparameters used across all models to ensure fair comparison and reproducibility.

In addition, a summary table has been included to list all evaluated models and their corresponding pre-trained weights, allowing readers to clearly identify the backbone configurations and initialization sources used in the experiments. These additions ensure that the experimental setup can be reproduced based on the information provided in the Methods section (Lines 250–261) and Table 1.

L216ff: What is the reasoning behind choosing these specific hyperparameter settings? Is there a loss curve that warrants that a training of 100 Epochs is enough?

Response: The hyperparameter settings were selected to balance training stability, convergence efficiency, and fair comparison across models, following configurations commonly used in recent remote sensing segmentation and classification studies. To justify the selected training lengths, training and validation curves have been added to the revised manuscript to document model convergence behavior.

These additions have been incorporated into the Results section (Lines 310–312; Lines 354-356; Figures 4 and 5).

Section 3.1: It is very good that the authors analyse and compare the performance of the different models for lake polygons, lake size, and elevation range using the MIoU. However, this could be complemented by an analysis of lake area, i.e. the ability by the different models to map the lake area as “completely as possible”. The analysis shown in Fig 6a already goes into this direction, where you can see that although, for example, DeepLab detects a lake as an entity, it fails to completely map the lake boundary as determined by the ground-truth data. I recommend a MioU analysis based on the total number of lake pixels detected by the different approaches.

Response: We acknowledge that this aspect was not clearly explained in the original manuscript. The ability to map lake areas as completely as possible is already evaluated through the pixel-based mIoU metric used in this study. In the revised manuscript, the definition of mIoU has been clarified by explicitly stating that it is computed over the entire validation dataset based on the total number of lake pixels, thereby directly quantifying the overlap between predicted and ground-truth lake areas.

This clarification has been added to the Methods section (Lines 265–269).

L270ff: To me, it was not immediately clear, why the authors chose to evaluate the performance of the models across elevation gradients. In the discussion, it turns out, that the authors associate different elevations with different environmental conditions (particularly vegetation cover and prevalence of snow). I agree, that the elevation gradient is a good proxy to model changing environmental conditions. However, I’d like a short (half-) sentence about that also in the results around L270 to avoid confusion.

Response: We agree that the rationale for evaluating model performance across elevation gradients should be stated more explicitly in the Results section. Elevation ranges are used here as a proxy for varying environmental conditions in alpine regions, such as differences in vegetation cover and snow prevalence.

In the revised manuscript, this rationale has been clarified by explicitly stating that stratified analyses across elevation ranges complement lake size–based evaluations, thereby avoiding potential confusion for readers (Lines 302–306).

Tables 4, 5 and 6: Please add the F1 score as an additional column.

Response: The F1 score has been added as an additional column to Tables 5, 6, and 7 in the revised manuscript.

Figure 6: While I think that the examples demonstrated here show very well the strengths of the proposed approach, for the reader it is difficult to generalize these strengths from only two samples. Consider showing 2-3 other examples for (a) and (b), respectively, as an Annex/Supplementary material to the paper to bolster your claim.

Response: We agree that additional examples help to better illustrate the generality of the observed model behavior. In the revised manuscript, Figure 6 has been expanded by adding four additional representative comparison examples.

Section 4.2: Several things need to be addressed in this discussion:

- First, the authors need to specify, for which area the dataset comparison was conducted. Is it the STPG region again?

  Response: The spatial scope of the dataset comparison has been explicitly clarified in the revised manuscript. It is now clearly stated that the comparison was conducted for the STPG region, with results discussed in relation to previous lake inventories for the same area (Lines 472-474).

- Second, when comparing their mapping results to those of existing datasets, the authors give an average size of lake area missed by the previous datasets. In terms of the relevance of very small lakes, it would also be good to know, how much of total lake area has been missed by the previous studies by including only lakes larger than a certain threshold in comparison to the newly proposed method. Similarly, it would be good to know the share of area of these very small lakes of total lake area. This way, the relevance of these small lakes would become clearer.

  Response: We agree that assessing the contribution of very small lakes to the total lake area is important. This has been addressed in the revised manuscript by explicitly quantifying the cumulative area of lakes omitted by previous inventories relative to the total lake area (Lines 496–500).

- Third, while it is plausible, that the proposed method detects more lakes than the comparison datasets due to their size threshold, also the possibility of overestimation of lakes needs to be discussed. You can use the false positive rates from the results section to make an estimate.

  Response: We agree that the possibility of lake overestimation should be discussed.

In the revised manuscript, this issue has been addressed by considering the occurrence of false positives in the segmentation results and the role of post-processing. The discussion clarifies that non-lake artifacts were limited and largely removed during refinement, supporting the interpretation that the observed differences primarily reflect improved detection of small lakes rather than systematic overestimation (Lines 506–513).

L419ff: As I understand it, the analysis provided here is supposed to demonstrate, how much more accurate the proposed lake classification approach is in comparison to drawing a 10 km buffer around a glaciated area and marking all lakes inside as “glacial” and all lakes outside as “non-glacial”. I see several issues with this approach:

- The approach (including the selection of the buffer distance) feels arbitrary. Of course, a simple buffer, particularly one of this size, will not be able to accurately discriminate lake types. Is there previous literature that uses this approach for lake classification?

  Response: The use of a 10 km glacier proximity threshold for lake type classification was already introduced in the Introduction as a commonly adopted practice in existing glacial lake inventories. To avoid potential confusion in the Discussion and to make the comparison framework clearer to readers, this point has now been explicitly reiterated in the Discussion section, with appropriate references provided (Lines 514-516).

- Also, the selection of the region is arbitrary. Why select a single glaciated mountain range and not analyse the entire STPG region or using the validation data?

  Response: The selection of a single glaciated mountain range was intended to provide a clear and focused illustration of the practical implications of the proximity-based classification approach. In the revised manuscript, this choice has been clarified by stating that a representative glaciated mountain sector in the eastern STPG was selected specifically for detailed visualization, rather than for statistical evaluation.

  This clarification has been incorporated into the revised manuscript (Lines 519–520).

- Without giving any number of correctly/incorrectly classified lakes by the two approaches (similar to the tables of the results section) the performance comparison is rather meaningless.

Response: The performance comparison has been strengthened by adding explicit numerical counts of correctly and incorrectly classified lakes. A new table has been included in the revised manuscript reporting TP, TN, FP, FN, and F1-score for the different classification approaches, enabling a quantitative and transparent comparison of classification performance (Lines 479–481; Table 8).

However, I agree that the confusion of glacial and non-glacial lakes particularly in close proximity of glaciers needs to be addressed and evaluated! Figure 4 shows a plausible pattern of lake classifications across the region, but how robust is the proposed method specifically in regions where both types of lakes co-occur? I can imagine an accuracy assessment similar to that in Table 4 based on a subset of lakes in very close proximity to glaciers (e.g. a 1km buffer around all glaciated areas as determined by the RGI). This would be something for the results section. The discussion then needs to pick-up on these results, and, if possible, compare the performance of the proposed method (regarding lake type classification) with the comparison datasets by Zhang (2024a,b).

Response: We agree that the robustness of lake type classification in near-glacier environments requires explicit evaluation. In the revised manuscript, this has been addressed by including a quantitative comparison within close glacier proximity using distance-based criteria (including a 1 km threshold) and the proposed ViT-based method, with results summarized in a new comparison table. The discussion now relates these results to existing inventories by Zhang et al. (2024a, b), allowing a direct assessment of classification performance in regions where glacial and non-glacial lakes spatially co-occur (Lines 483–485; Table 8).

Discussion section in general: Are there significant differences in computational effort between the compared DL models? If so, do the authors think the increase of accuracy is worth the additional effort?

Response: We agree that computational effort is an important consideration. In the revised manuscript, this has been addressed by noting that the evaluated models do not exhibit order-of-magnitude differences in computational cost and that, for offline inventory-scale mapping, accuracy and robustness are prioritized over computational efficiency (Lines 465-470).

Technical corrections

L15-16: Add commas to sentence to enhance readability

L17-18: Suggestion: “challenges for conventional identification methods”

L26: “provided **a** more accurate lake type classification”

L80: The is a space too much here

L86: Remove full stop before citation

L97: add “the” or “a” before ViT architecture

Fig 6: Format figure caption consistently.

Fig 7: Please add scales to all of the images, and some kind of indication, where the area is located (e.g. map inset or geographic coordinates).

L394f and Table 7: Inconsistency in the citation of the Zhang (2024) papers. Please either use (2024 a/b) OR G. Zhang/T. Zhang (whichever fits best to the journal’s preferred citation style).

L459: lower case o in “Overestimation”

Response: We appreciate these detailed editorial and formatting suggestions. All listed issues have been addressed in the revised manuscript, including minor grammatical corrections, terminology refinements, and consistency improvements in citations, capitalization, punctuation, and figure formatting. Scales and location information have been added to figures where requested.

L164-165: The links provided refer to the data portals but not the datasets. Please provide links to the respective data catalogue entries of the platforms. If these links are to long, consider a scientific citation of the original data.

L215: Please provide direct links to the models and datasets instead of just to the platform.

Response: In the revised manuscript, the general portal links have been replaced with direct links to the specific catalogue entries of the referenced datasets and models (Lines 177-181; Table 1).

L364: I wouldn’t call the use of various spectral bands and calculated indices thereof “multisource remote sensing data”, when all of the products come from a single system (Sentinel-2).

Response: Thank you for pointing this out. The term “multisource remote sensing data” has been corrected in the revised manuscript to avoid ambiguity, as all inputs are derived from

Sentinel-2. The wording has been revised to “multi-dimensional remote sensing features” to more accurately reflect the use of multiple spectral bands and derived indices from a single sensor.

RC2

The authors present a new method for automatic mapping and classification of high-mountain/glacial lakes applied to the Third Pole region. The manuscript is generally well composed and the method presented represents a valuable addition to the approaches applied so far. The contribution is of high relevance since knowledge on lake distribution, especially in climate-change affected mountains, is essential for hazard management and mitigation.

While the method and results are well presented and the discussion and conclusion are largely convincing, the manuscript suffers severely from a poor application of the terminology and definition of glacial and non-glacial lakes. This has little effect on the lake detection itself but huge implication on the lake classification and the results and comparison in general. With respect to the potential relevance of the produced dataset for hazard management, this issue needs to be resolved. Otherwise, despite its technological performance, the dataset will be of little use.

To conclude, I think this manuscript requires a revision with respect to the application of the right terms for the objects in focus. In addition, more attention should be paid to the introduction of the comparative database to improve clarity. These revisions require moderate effort, will not affect the geometry of the lakes dataset but surely will change the classification and the discussion. This will improve the quality of the study and ensure comparability and a wider application of the dataset in the intended way.

Response: We thank you for the positive assessment of our technical framework and for highlighting the relevance of this work in the context of hazard management. We also appreciate the critical feedback regarding the terminology and definition of “glacial” and “non-glacial” lakes. While the deep learning model demonstrates strong technical performance, we recognize that the scientific value of the resulting dataset depends on a rigorous and geomorphologically sound classification scheme. Accordingly, the terminology has been carefully revised throughout the manuscript to ensure conceptual clarity and consistency.

Detailed point-by-point responses to your specific comments are provided below.

Terminology: The authors need to reconsider the definition of glacial and non-glacial lakes. In the manuscript a variety of terms are applied starting with the term periglacial lakes in the title and introduction (and not more afterwards) than glacier lakes, montane lakes and non-glacier lakes. The authors mention to follow the classification by Yao et al. (2018) but a

detailed definition of the terminology is absolutely required. This will influence the results and interpretation. For additional clarification I suggest fundamental review papers on the terminology for example by Carrivick and Tweed (2013) [DOI: 10.1016/j.quascirev.2013.07.028]

Furthermore, the title is confusing. Despite the use of the term “periglacial lake”, I also don’t know what “global attention” is signifying in this context. Please reconsider a more appropriate title.

L39ff - You should provide a better definition of non-glacial lakes. The reference to “thermodynamic processes" is not enough from a geomorphological perspective since this is a too broad term from physics. The term periglacial lakes is not commonly used, since the formation is not linked to periglacial processes (involving ground ice and freeze-thaw). Using periglacial lakes with respect to the location of the lake should be avoided due to the misleading connotation of the term periglacial here.

L58ff – same issue as above…

Response: We sincerely appreciate the guidance on terminology. We have carefully studied the suggested literature by Carrivick and Tweed (2013) and acknowledge that our initial use of “periglacial lakes” as a single category was not sufficiently rigorous.

In the revised manuscript (Lines 31-51), we clarify that our study focuses on lakes in periglacial environments. Within this environmental setting, a binary classification has been implemented.

Glacial lakes: follow the definition of Yao et al. (2018), referring to natural water bodies mainly supplied by modern glacial meltwater or formed in depressions of glacial moraines. This definition was adopted because the manual interpretation and data labeling in this study were conducted strictly according to these established criteria, ensuring full consistency between the methodology and the scientific definitions.

Non-glacial lakes: include all other water bodies within the periglacial study area that do not meet the above criteria.

Furthermore, we agree that the term “global attention” may be misinterpreted as referring to global interest rather than a modeling concept. To avoid ambiguity, this term has been replaced with “global context” throughout the manuscript to better reflect the long-range contextual information captured by the Transformer architecture. In addition, acknowledging that the term “empowers” may be overly strong, it has been replaced with the more precise term “enhancing.” Accordingly, the title has been revised to:

“Enhancing Lake Identification in Periglacial Environments by Leveraging the Global

Context of Transformers."

To illustrate this, one must investigate chapter 3.3: In the STPG region most of the non-glacial lakes identified and depicted in Fig. 4 are indeed glacial lakes, according to most classification schemes, because they have been formed by glacial erosion. Many are found in cirques that have been sculpted by glaciers (e.g. in the area around 29°,11.441' N/95°33,340'E). The only difference is that they are located in catchments without current glaciers, thus they have been formed by glacier action in the past. Your terminology should therefore not only include a geomorphological and topographic definition, but also a temporal one (see for example Buckel et al. (2018)). Non-glacial, from my perspective would be restricted to lakes formed by landslides/debris flow dams or of volcanic origin. Lakes purely formed by excessive precipitation are vary rare in mountainous regions from my perspective.

My suggestion would be to either add a temporal aspect to your definition (Holocene, historic glacial lake) or to only to focus on ice-contact or near-glacier lakes (which would involve a distance-based definition).

This terminological uncertainty should be resolved and then considered in the discussion of the distance-based method. Your comment may of course be valid for some applications esp. natural hazards assessment (e.g. GLOF), but some of the argumentation is lost when the terminology is better defined and applied. In this respect authors need to consider that the distance-based method is justified here, assuring that there is a glacier upslope of the lake.

Response: We appreciate the insightful discussion regarding the geomorphological origin and temporal dimension of lake formation. We acknowledge that many lakes located in cirques and other relict glacial landforms were historically shaped by glacial erosion and could be considered glacial in a long-term geomorphological sense.

In this study, lake classification follows the framework of Yao et al. (2018), which is explicitly based on modern glacial influence and present-day hydrological processes, rather than on the long-term geomorphological origin of lake basins. Our initial manuscript did not state this premise with sufficient clarity. In the revised manuscript, this has now been explicitly clarified: glacial lakes are defined as water bodies directly coupled to ongoing glacier dynamics through modern meltwater supply or moraine-dammed settings, whereas lakes formed in relict glacial landforms without current glacier influence are categorized as non-glacial under this process-based definition (Lines 36-40).

The implications of this process-based definition for proximity-based classification approaches are now made clearer in the Discussion, building on the commonly adopted 10 km glacier proximity threshold already introduced in the Introduction and reiterated in the Discussion with appropriate references (Lines 514–516).

Some minor comments:

L36– exchange the term “montane” with “alpine/high-alpine” – montane refers a biogeographic altitudinal zone usually at intermediate altitudes. (throughout the manuscript!!)

Response: Thank you for the correction. We agree that “alpine” more accurately reflects the high-altitude context of this study. Accordingly, the term “montane” has been replaced with “alpine” throughout the manuscript, including the revised title.

L251ff – You compare the result to other approaches (CNN, UNet, DeepLapv3+), but you don’t mention that you applied these methods as well. How was this comparison done? Did you use existing data from other studies? This need to be mentioned in the methods section (e.g. 2.5) and reference in Table 1.

L282 – Ch 3.2 – Similar to the comment above – You compare your classification results with two other CNN approaches (EfficientNet, ResNet). How was this done? Again no mentioning in the methods before.

Response: The comparison with CNN-based approaches was conducted by implementing and training all comparative models within this study, rather than relying on results from previous publications. UNet and DeepLabv3+ were trained for lake outline segmentation, and ResNet and EfficientNet were trained for lake type classification, all using the same dataset and a unified training strategy to ensure fair comparison. This has now been explicitly described in the Methods section (Lines 250-260), including the training setup and hyperparameter configuration, and the corresponding models and references have been clearly documented in Table 1.

L269 – Table 2 (and same for table 3): The tables hold the category “all”. What does this mean? Are these the mapped lakes? I suggest renaming this class for better clarity.

Response: We agree that the category “All” was ambiguous. In the revised manuscript, this category has been renamed to “Final inventory” in Tables 3 and 4 to clearly indicate that it refers to the reference lake inventory used for evaluation.

L291 – Add explanation for TP, FP, TN, FN in the table caption.

Response: The definitions of TP, FP, TN, and FN have been explicitly provided in the Methods section together with the corresponding evaluation metrics. This has been clarified

to ensure that the meaning of these terms used in the tables is clear and unambiguous.

L329 – Exchange “The proposed framework” with a more precise description excluding the CNN/alternative methods. Like: ViT-based methods…

Response: The wording has been revised by replacing “the proposed framework” with “the ViT-based identification framework” to more precisely refer to the method under discussion.

L395ff – Chap 4.2 – please add the a, b to the Zhang references throughout the chapter to better differentiate between the publications.

Response: The references have been updated as requested to ensure consistent citation throughout the chapter.