

RESPONSE TO REVIEWER 2

SETP_GLI: An annual 10–30 m glacial lake inventory for the southeastern Tibetan Plateau from 1990 to 2025

We thank the reviewer for the careful assessment of the GLA-RCNN architecture, training data, validation design, sensor transition, uncertainty treatment, and presentation. We have made targeted textual revisions in the manuscript based on the current dataset and processing workflow. The revised text clarifies the model configuration, contemporaneous label generation, the 30 m Landsat temporal baseline, pixel-level cloud-gap compositing, the retained 10 m Sentinel-2 companion layers, validation limitations, image-selection records, and uncertainty sources. Figure-file updates are identified in the relevant change summaries. The response also distinguishes the completed 2013–2025 Landsat 7/8/9 processing from the planned target-year pixel-level supplementation and re-extraction of the 1990–2012 Landsat record. This work is expected to improve continuity for variable glacial lakes and reduce the need for manual corrections; the corresponding statistics will be recalculated after reprocessing and quality review.

Major comment 1 - GLA-RCNN architecture description

Reviewer comment

GLA-RCNN Architecture Description is unclear. Please provide a detailed architecture diagram showing where CBAM modules are inserted, specify the reduction ratio used in channel attention, and give the spatial-attention kernel size.

Response

We clarified the architecture in Section 4.2. CBAM is inserted after ResNet-152 Conv4_x and Conv5_x, and after the FPN outputs P2–P5 and the pooled P6 level. The channel-attention reduction ratio is $r = 16$ and the spatial-attention convolution uses a 7×7 kernel. The attention maps sequentially reweight each feature map by elementwise multiplication. The architecture schematic will be updated accordingly in the final figure set.

Changes made in the manuscript

- Section 4.2: added the CBAM insertion points, reduction ratio, spatial kernel size, and multiplicative feature-reweighting description. Figure 3: schematic clarification is reserved for the final figure replacement.

Major comment 2 - Training data generation

Reviewer comment

The methodology for the 1991 training node is ambiguous because Wang et al. (2020) provides inventories for 1990 and 2018. Clarify how contemporaneous imagery was used for 1991 digitization, who performed the digitization, the quality-control protocol, and inter-operator consistency metrics.

Response

Thank you for this comment. The 1991 training labels were manually delineated from contemporaneous 1991 imagery within the study area. The public inventory of Wang et al. (2020) was used to locate candidate lakes, rather than to supply their 1991 boundaries. One operator examined the candidate lakes individually against the 1991 imagery and delineated their boundaries according to the visible lake extent at that time. Thus, the labels were not copied directly from the 1990 or 2018 inventory. The resulting labels were then reviewed by two other operators to check lake identification and boundary correspondence with the same imagery. Quality control therefore consisted of individual image-based delineation followed by review by two additional operators. A separate quantitative inter-operator agreement statistic was not recorded, and we do not claim a quantified level of agreement.

Changes made in the manuscript

- Section 4.3: clarified the source and timing of the 1991 imagery, the role of Wang et al. (2020), and the three-person cross-check procedure. Section 8.4: noted that inter-operator agreement was qualitatively reviewed but not quantified.

Major comment 3 - Sensor transition handling

Reviewer comment

Explain whether separate models were used for 1990–2012, 2013–2015, and 2016–2025, how predictions were harmonized, and how performance differences across resolutions were addressed.

Response

Thank you for this comment. The three resolutions did not correspond to three independently designed models. A common GLA-RCNN architecture was used, with adjustments to the sensor-specific input configuration, including the available spectral bands and input-channel number. Resolution-dependent performance was evaluated separately for the 30, 15, and 10 m inputs, as reported in Table 4.

To improve long-term temporal consistency, we reprocessed the 2013–2025 observations using Landsat 7, 8, and 9 imagery on a common 30 m grid, extending the 30 m Landsat record beyond 2012. We will also apply target-year pixel-level supplementation and re-extraction to the 1990–2012 Landsat record to improve the continuity of observations of glacial lakes with variable extents. The 15 m Landsat results for 2013–2015 and the 10 m Sentinel-2 results for 2016–2025 were retained as complementary higher-resolution datasets rather than directly concatenated with the 30 m record. These overlapping products preserve detailed information on small lakes and complex shorelines and support cross-checking of lake presence and boundary delineation.

The revised framework therefore distinguishes the 30 m temporal baseline from the higher-resolution companion products. Maintaining a common mapping resolution reduces resolution-driven

discontinuities, although differences in spectral response, acquisition conditions, and boundary detectability remain. The common $\geq 0.01 \text{ km}^2$ analysis threshold provides additional screening but is not interpreted as eliminating these residual differences.

Changes made in the manuscript

- Section 4.3: clarified the distinction between the 30 m Landsat temporal baseline, the retained 10 m Sentinel-2 companion layers, the original 15 m validation benchmark, and the pixel-level compositing rule; native-grid outputs are not treated as geometrically equivalent boundaries.

Major comment 4 - Validation circularity

Reviewer comment

Hi-MAG and Wang et al. (2020) assisted label generation and were also used for accuracy assessment. This dual use can introduce circularity bias and penalize newly detected small lakes absent from older 30 m references.

Response

We agree that the internal held-out chip metrics and the external inventory comparisons answer different questions. During label preparation, Hi-MAG and Wang et al. (2020) were used only to locate and cross-check candidate lake positions; the final boundaries were delineated from contemporaneous imagery rather than copied from the reference polygons. The labels also included observable lakes absent from the public inventories, including lakes that were not necessarily micro-lakes. The 8:1:1 chip split provides held-out image-chip evaluation, but the public inventories also informed candidate selection, so strict independence cannot be claimed for the external comparisons. Shared reference information may increase agreement for common lakes, whereas valid lakes absent from older 30 m inventories may be counted as unmatched detections and reduce apparent precision. Because no independent high-resolution inventory covering the full SETP and the complete study period was available, we retain these comparisons as consistency assessments and explicitly acknowledge this limitation in the manuscript. Accordingly, these external comparisons are described as cross-inventory consistency assessments, not as independent estimates of absolute mapping accuracy.

Changes made in the manuscript

- Sections 4.5 and 8.4: added an explicit limitation concerning reference overlap, circularity, and the resolution-related penalty applied to newly detected small lakes.

Major comment 5 - Ground truth quality for small lakes

Reviewer comment

Google Earth imagery has heterogeneous dates and resolutions. Provide dates, resolution, quality information, and discuss temporal misalignment with Landsat and Sentinel-2 acquisitions.

Response

Thank you for this comment. We agree that the Google Earth imagery used in the small-lake checks has heterogeneous acquisition dates and effective spatial resolutions across the SETP. Its primary purpose was to verify whether lakes smaller than 0.01 km² were detected and whether their general spatial locations and shoreline shapes were plausible, rather than to provide temporally exact, pixel-level ground truth. Exact nominal spatial-resolution metadata were not consistently available for all Google Earth scenes, so unsupported resolution values were not assigned; the imagery is treated as a qualitative high-resolution reference rather than temporally exact pixel-level ground truth.

Changes made in the manuscript

- Section 5.2: added the temporal and spatial heterogeneity caveat for Google Earth reference imagery and its implications for small-lake checks. Figure 9/10: metadata and legibility updates are reserved for final figure replacement.

Major comment 6 - Uncertainty quantification**Reviewer comment**

The boundary-pixel method does not account for georeferencing errors, seasonal water-level fluctuations, or shadow-induced boundary shifts, especially in early Landsat scenes. Additional uncertainty sources should be discussed.

Response

Thank you for this comment. The boundary-pixel method of Hanshaw and Bookhagen (2014) was used specifically to quantify the area uncertainty caused by mixed pixels and the pixel-scale irregularity of lake boundaries. It does not represent a complete uncertainty budget.

We agree that additional sources of uncertainty, including residual georeferencing errors in early Landsat imagery, seasonal water-level fluctuations, snow and ice cover, terrain-shadow effects, and debris-covered shorelines, were not independently quantified. We have therefore clarified the scope of the boundary-pixel estimate and expanded Section 8.4 to discuss these additional uncertainty sources and their potential effects on lake boundaries and areas. The reported values are therefore termed boundary-related area uncertainty, not total mapping uncertainty.

Changes made in the manuscript

- Section 8.4: expanded the limitations of the boundary-pixel uncertainty estimate to include georegistration, seasonal, shadow, snow/ice, and debris-related effects.

Major comment 7 - Temporal consistency and sensor-specific thresholds

Reviewer comment

Temporal Consistency Issues. Lines 102-119 mentioned the primary window for image selection are Sept-Nov, backup August, then Dec-Feb. This introduces seasonal variability that could obscure long-term trends. Lakes in December may be partially frozen; August lakes may be at peak melt. Can quantify the proportion of lakes from each month and analyze whether there's a systematic bias in lake area from different months. Lines 226-233 say sensor-adaptive thresholds of 0.001 km² (10m), 0.002025 km² (15m), and 0.0054 km² (30m). This means the minimum detectable lake size changes over time. Lines 473-485: There's a gap between Landsat and Sentinel-2 coverage. How was 2014-2015 handled? Was Landsat 8 used exclusively? What about 2016 transition year? The author can provide detailed per-year information on primary sensor used, the number of scenes and cloud cover statistics and any data gaps and how they were filled.

Response

Thank you for this detailed comment. We revised the image-selection and temporal-consistency procedures and clarified the sensor-transition treatment.

The preferred observation window is September–November, when seasonal snow and cloud effects are generally reduced and lake boundaries are comparatively stable. Other observations within the same target year, including adjacent-month acquisitions, were used only when suitable September–November coverage was unavailable. For the reprocessed 2013–2025 Landsat baseline, clear pixels from the priority observation were retained, and subsequent suitable observations were used only to fill cloud- or invalid-value gaps at the same locations; valid pixels were never overwritten. This pixel-level compositing reduces seasonal discontinuities caused by image availability, although it cannot completely remove effects from seasonal water-level changes, freezing, or snow cover. We expanded the supplementary annual image records to report acquisition dates and months, sensors, scene or tile information, cloud or data-gap status, and the source of local gap filling, together with the monthly observation distribution needed to assess potential seasonal bias in mapped lake area. A separate numerical sensitivity test quantifying area bias between preferred- and fallback-month observations was not performed in this revision; this remains a limitation when interpreting individual lake areas. No adjacent-year imagery was used as a direct substitute for a target-year boundary. If suitable target-year coverage remained unavailable, the annual observation was treated as unavailable or uncertain. This approach reduces the possibility that cross-year filling could obscure genuine interannual changes.

We also clarified that the 0.001, 0.002025, and 0.0054 km² thresholds are sensor-specific minimum mapping units for 10, 15, and 30 m imagery, respectively. They are practical screening thresholds and do not imply equivalent detection capability across the full time series. The uniform ≥ 0.01 km² threshold is applied only for aggregate trend screening and reduces, but does not eliminate, cross-sensor differences.

To improve temporal comparability, we reprocessed the 2013–2025 annual baseline using Landsat 7, Landsat 8, and Landsat 9 imagery on a common 30 m grid. Thus, the 2014–2015 and 2016 baseline layers were generated using the same 30 m Landsat processing framework, with pixel-level supplementation where necessary; Landsat 8 was not treated as the sole source for all years. The 15 m Landsat 8 results for 2013–2015 and the 10 m Sentinel-2 results for 2016–2025 were retained as higher-

resolution companion products. Sentinel-2 therefore supplements the 30 m baseline after 2016 rather than replacing it in the primary long-term record.

We will also extend pixel-level supplementation and re-extraction to the 1990–2012 Landsat record, using suitable observations from each target year to fill cloud-obscured or invalid pixels while retaining valid observations. This is intended to improve the continuity of observations of glacial lakes with variable extents by reducing gaps caused by incomplete image coverage. Improved image coverage is expected to reduce the need for manual corrections; the resulting correction counts and percentages will be recalculated after reprocessing and quality review. The manual-correction statistics currently reported refer to the existing native-resolution inventory.

Changes made in the manuscript

- Sections 3.2 and 4.3: clarified the preferred September–November acquisition window, adjacent-month gap handling, annual multi-acquisition mosaics, pixel-level Landsat cloud-gap supplementation, and the common 30 m Landsat 7/8/9 baseline for 2013–2025. Section 5.3: clarified the observation window used when interpreting cross-dataset consistency metrics. Section 8.2: clarified the residual seasonal and cross-sensor limitations. Supplementary Table S1: expanded the annual scene/date and source-image records used for classification, including the information needed to assess acquisition-month distribution and gap filling.

Major comment 8 - Comparison-model training fairness

Reviewer comment

Were Mask R-CNN, U-Net, and DeepLab V3 optimized for this task? Specify learning rate, epochs, augmentation, hyperparameter tuning, and whether the comparison used equal training resources.

Response

The comparison models used the same 1991 source-chip split (seed 42; 8:1:1), 256 × 256 chips, normalization, AdamW, learning rate 1×10^{-4} , weight decay 1×10^{-4} , a maximum of 30 epochs, and best-validation-checkpoint selection. Random 90-degree rotations and horizontal or vertical flips were applied only to the training subset; validation and test subsets were loaded without augmentation. No model received additional data or a larger epoch budget, and no model-specific hyperparameter search was performed. Detection and semantic-segmentation models used their memory-compatible batch sizes (2 and 8, respectively, in the comparison configuration), which is now acknowledged as a practical implementation difference. We therefore interpret this experiment as a controlled baseline comparison under a common optimization budget, rather than as an exhaustive comparison of individually optimized architectures.

Changes made in the manuscript

- Section 4.3: added the common comparison-model training settings, checkpoint rule, and the batch-size qualification.

Minor comment 1 - CBAM integration with FPN

Reviewer comment

How were the CBAM outputs integrated with the FPN features? Provide a precise explanation and schematic.

Response

For each FPN output, CBAM first generated channel-attention and spatial-attention maps. The feature map was multiplied element by element by the channel-attention map and then by the spatial-attention map before being passed to the subsequent FPN or prediction operation. The revised text also identifies P2–P5 and the pooled P6 output. The schematic will be updated for readability in the final figure set.

Changes made in the manuscript

- Section 4.2: added the sequential multiplicative FPN integration description and the pooled output. Figure 3: schematic update is reserved for final figure replacement.

Minor comment 2 - Sensor-specific input channels

Reviewer comment

Clarify how variable input dimensions are handled between Sentinel-2 and Landsat and discuss potential spectral inconsistencies.

Response

The input configuration remains sensor-specific. The 30 m Landsat 7/8/9 inputs contain seven channels (five harmonized optical bands plus NDWI and slope), whereas the Sentinel-2 inputs contain six (four optical bands plus NDWI and slope) because a 10 m SWIR band is unavailable. Bands are normalized before stacking with NDWI and slope. The prepared image chips determine the channel count through the `in_channels`: auto option, and that count remains fixed during each training and inference run. The 30 m Landsat and 10 m Sentinel-2 feature spaces are therefore not assumed to be interchangeable.

Changes made in the manuscript

- Sections 4.1 and 4.2: clarified automatic channel configuration, the six- versus seven-channel inputs, normalization, and the resulting spectral non-equivalence.

Minor comment 3 - Intuitive loss-function explanation

Reviewer comment

The loss description is technically dense. Add a brief intuitive explanation of each loss term.

Response

We added a short explanation: classification term distinguishes lake instances from background, box-regression terms align candidate extents, and mask loss refines pixel-level boundaries.

Changes made in the manuscript

- Section 4.2: added the intuitive explanation immediately after the mathematical loss definition.

Minor comment 4 - Discussion focus

Reviewer comment

The Discussion repeats the Introduction. Focus it on implications of the new dataset rather than repeating its features.

Response

We condensed the repeated introductory description and refocused the Discussion on the interpretation of sensor-dependent trends, uncertainty, validation scope, applications, and limitations.

Changes made in the manuscript

- Section 8.1: revised the opening discussion to emphasize dataset implications and interpretation rather than repeat the Introduction.

Minor comment 5 - Figure 2 terminology

Reviewer comment

Correct 'STREM' to 'SRTM' in Figure 2.

Response

The text uses the correct SRTM terminology. The figure label will be corrected in the final figure file.

Changes made in the manuscript

- Figure 2: spelling correction reserved for the final figure replacement; related body text remains consistent with SRTM.

Minor comment 6 - Figure 3 readability

Reviewer comment

Ensure all text can be read clearly in every panel of Figure 3.

Response

We will enlarge and simplify the panel text and labels in the final architecture schematic. The manuscript text now supplies the exact module locations and parameters independently of the figure.

Changes made in the manuscript

- Section 4.2: architecture details added in text. Figure 3: legibility update reserved for final figure replacement.

Minor comment 7 - Figure 10 raw and MMU trends

Reviewer comment

What do 'raw' and 'MMU' trends represent? Explain this in the caption or main text.

Response

We clarified the terminology. The full-inventory trend is the annual inventory after the sensor-specific mapping thresholds. The MMU trend is the subset retained after applying the uniform ≥ 0.01 km² analysis threshold. The latter reduces, but does not eliminate, cross-sensor differences.

Changes made in the manuscript

- Annual-trend caption and Section 7: defined the full-inventory and MMU series and qualified their use for cross-period comparison.