Estimating Glacial Lake Trends using Historically Guided Segmentation Models

Anthony Ortiz*, Weiyushi Tian*, Tenzing C. Sherpa, Finu Shrestha, Mir Matin, Rahul Dodhia, Juan M. Lavista Ferres, and Kris Sankaran

***Abstract*—We compare several approaches to segmenting glacial lakes in the Hindu Kush Himalayas in order to support glacial lake area monitoring. More automatic monitoring could support risk assessments of Glacial Lake Outburst Floods (GLOF), a type of natural hazard that poses a risk to communities and infrastructure living in valleys below glacial lakes. We evaluate several approaches to incorporate labels from a 2015 survey using Landsat 7 imagery to guide segmentation on newer higher resolution satellite images like Sentinel 2 and Bing Maps imagery, comparing them also to approaches that do not use this form of weak prior. We find that a guided-version of U-Net and a properly initialized form of morphological snakes are most effective for these two datasets, respectively, each providing between an 8 - 10% IoU improvement over a standard U-Net. An error analysis highlights the strengths and limitations of each approach. We design visualizations to support discovery of lakes of potential concern, including an interactive exploratory interface. All code supporting our study are released in public repositories - Models, Visualizations, Experiments.**

***Index Terms*—Glacial Lakes, Glacier Monitoring, Deep learning, Morphological Snakes, Climate Change**

THIS work focuses on monitoring of glacial lakes using historically guided semantic segmentation models and satellite imagery. Glaciers all over the world are currently thinning and retreating at a remarkable rate in the recent decades as a result of the changing climate [6]. The melting of glaciers often accumulate to form *glacial lakes* between the frontal moraine and the retreating glacier or on the surface of the lower section of the glacier [8]. These lakes are dammed by moraine complexes which are often unstable and have a potential to breach. A *Glacial Lake Outburst Flood (GLOF)* event occurs when a dam holding a glacial lake collapses. The resulting rush of water and accumulated debris can cause significant damage to nearby communities destroying human lives and infrastructure. GLOFs have led to more than 12,000 deaths and destroyed roads, bridges, hydroelectric developments, and entire villages [4, 20]. Accurate delineation and monitoring of glacial lakes is required for determining the risks of such GLOF events [8] and developing GLOF mitigation strategies.

Different organizations around the world, including The International Centre for Integrated Mountain Development (ICIMOD) have been working on creating glacial lake inventories [8, 12, 23, 3, 14, 18, 25, 11]. Until now, most of these inventories have been opting for manual methods of delineating glacial lakes using optical or SAR imagery [8, 12, 23, 3, 15]. There have also been several studies who have applied semi-automated approaches for segmenting glacial lakes [18, 25, 11, 22, 9]. For example, [11] generated superpixels on satellite imagery using an object-based segmentation model from the eCognition software application [16]. These superpixels on the surface of glacial lakes were then manually selected and the rest were removed. For the Hindu Kush Himalayan (HKH) region, ICIMOD developed an inventory of 3,624 glacial lakes covering the three major basins from Nepal, Tibet, and India, of which 47 were classified as at high-risk for GLOF events, based on the lake and dam characteristics, source glacier activity, and surrounding morphology [8, 11]. In this study, Landsat 7 remote sensing imagery was segmented using a Normalized Difference Water Index (NDWI) as the input to an object-based segmentation model, also from eCognition. The segments other than glacial lake were manually selected. The missing lakes or lake segments covered with clouds or shadows were manually added and refined using image scenes from a different date. Once the final glacial lakes were prepared, their area was computed. Comparing the area and expansion rate of the lakes across years provides one indication of GLOF risk.

The glacial lake inventories are only infrequently updated. This is because (a) identifying lakes and delineating their boundaries is a time consuming and laborious process, and (b) Landsat imagery has 30 meter resolution, making smaller changes in glacial lake area impossible to distinguish. Since updates are infrequent, it is difficult to observe trends in lake properties. New lakes may have entered into a high-risk state, while those thought to be high-risk may have stabilized. While there are only 47 lakes currently classified as high-risk, there is a much larger number that may need to be screened periodically.

Researchers have shown interest on developing algorithms to fully automate glacial lake mapping and monitoring. Li and Sheng proposed an automated glacial lake delineation algorithm based on hierarchical image segmentation and digital elevation maps [10]. In this algorithm, each glacial lake is delineated using segmentation value, and the topographic features derived from digital elevation models (DEMs) are used to separate mountain shadows from glacial lakes. However, this method is highly sensitive to changes on image conditions. More recently, researchers started exploring the use of deep learning models to outline glacial lakes [19].

A. Ortiz and W. Tian contributed equally as first authors to all academic and professional effects, and their names can be legitimately swapped in their respective publication lists

A. Ortiz, Rahul Dodhia, and J.M. Lavista are with the AI for Good Research Lab, Microsoft.

E-mail: anthony.ortiz@microsoft.com

W. Tian and K. Sankaran are with the University of Wisconsin-Madison.

T. Sherpa, F. Shrestha, and M. Matin are with the International Centre for Integrated Mountain Development (ICIMOD)

Under Submission

Qayyum et al. [19] trained a U-Net model [21] to segment glacial lakes using PlanetScope imagery. Fang Chen also used U-Net to segment supra-glacial lakes in the Himalayas using SAR images [5], while Wu et al. [27] trained a U-Net variant using a combination of optical and SAR imagery as input.

In this work, we evaluate several approaches to incorporate historical glacial lake information and guide segmentation on newer and higher resolution satellite and aerial imagery from Sentinel 2 and Bing Maps, comparing them to approaches that do not use this form of weak prior like the ones previously described. We find that a guided-version of U-Net and a properly initialized form of morphological snakes are most effective for these two datasets, respectively, each providing between an 8 - 10% IoU improvement over a standard U-Net used on previous work. We also develop an interactive visualization tool to create a solution to monitor glacial lakes at very high resolution.

Our contribution is three-fold:

- We incorporate historical glacial lakes information at lower resolution to guide segmentation on higher resolution Sentinel 2 and Bing Maps imagery.
- We evaluate several approaches and show that our proposed guided segmentation greatly outperform other glacial lake mapping approaches.
- We design visualizations and an interactive exploratory interface to facilitate the discovery glacial lakes of potential risk.

I. Methods

The general lake monitoring problem is a multitemporal segmentation problem with images $x_{it} \in \mathbb{R}^{W \times H \times D}$ of lake i across N_i times t. pixel-level annotations of the lake at time t are encoded in $y_{it} \in \{0,1\}^{W \times H}$. The segmentation task is to supply a new annotation $\hat{y}_{it^*}$ given a new observation x_{it^*}.

In our application, we encounter additional constraints. First, label availability varies across timepoints t. We have access to a large number of labels y_{it} for earlier timepoints, corresponding to the 2015 comprehensive survey conducted by ICIMOD [11]. However, very few labels are available at more recent timepoints, and for most t, no labels are available. The second constraint is that, for some newer, higher-resolution sensors, only recent imagery x_{it} available. This nonuniformity in label availability over time complicates the use of standard multitemporal segmentation strategies, precluding application of methods like ConvLSTM for segmentation of earth observation time series [28]. A challenge of this study is to understand how to effectively use historical annotations within this context. We next discuss five strategies drawing from across the image processing literature, making use of weak supervision to varying degrees.

A. U-Net and Historically Guided U-Net

U-Net is a standard convolutional neural network architecture recently used for the glacial lake segmentation problem [21, 19, 13, 7]. In its standard form, the U-Net architecture does not directly make use of historical annotation. The method learns a mapping between x_{it} and matched labels y_{it} which can be applied to recent images x_*, but without attempting to match x_* to past labels.

We additional consider a historically guided version of U-Net which adds a new channel to x_{it} including a partially ablated label $A(y_{it'})$ from a time $t' \leq t$. The ablation is needed to prevent the model from simply "copying" the past label to the current timepoint, a strategy that may work artificially well for unchanging lakes, but which fails to uncover the changes in lake boundary that are of interest. We set $A(y_{it})$ to be a reverse buffered version of polygons induced by y_{it}, shrunken equally in all directions until the covered area is reduced to 40% of the original area.

B. Morphological Snakes

Active contour methods frame image segmentation as a variational problem [17, 2]. Each image defines an energy functional, and the boundaries that define a segmentation can be associated with an energy. Curves that closely follow edges and bound regions of similar pixels will have low energy, while curves that cross edges or contain patches with different image statistics will have high energy. In this way, a segmentation can be found for each image x_i by identifying a set of boundary curves C_i with low overall energy.

Formally, the morphological snakes algorithm parameterizes an evolving sequence of two-dimensional segmentation boundaries using a sequence of mappings $\varphi_\tau : \mathbb{R}^2 \to \mathbb{R}$. The 0-level set of φ_τ defines the segmentation at iteration τ. That is pixels laying in $\{x : \varphi_\tau(x) < 0\}$ are classified as lake at iteration τ. Note that this level-set formulation allows segmentation of disconnected regions. To find a segmentation with minimal energy, the level set φ_τ is evolved by iteratively applying three operations — dilation, erosion, or curvature flow — which grow, shrink, and smooth out the segmentation boundary, respectively. The strength of these operations is determined by local image statistics in order to more closely follow object boundaries. We refer to [2] for implementation details; our experiments use the open-source library [24]. This method requires an initialization φ_0 to define preliminary image statistics that should be reflected by the final segmented object. We initialize φ_0 with the level set function derived from the same ablated labels $A(y_{it'})$ used in the historically guided U-Net.

C. Deep Level-Set Evolution

We also apply the DELSE algorithm, a level set method that is able to learn update rules based on deep, data-specific features [26]. Like morphological snakes, this algorithm takes advantage of a coarse, approximate labeling to guide a full segmentation. It also defines a sequence of level set functions $\varphi_\tau(x)$ in pixel space that is negative if and only if it lies inside the class of interest.

In addition to imagery x_i, DELSE expects extreme-point supervision for each polygon. These are the four points defining a bounding box of the polygon label. These points are rasterized into an image of the same dimension as the image of interest x; specifically, a Gaussian-blur resampling is applied. Call the rasterized extreme-point labels w. The

DELSE algorithm initializes a level set $\varphi_0 = f_{\theta_1}(x_i, w_i)$ approximating the locations of polygons. Then, it evolves this surface $\varphi_{\tau+1} = \varphi_\tau + U_{\theta_2}(x_i, \varphi_\tau)$ for T steps using an update function U_{θ_2}. Note that, due to the vanishing gradient problem, T is typically chosen in the range 3 - 5. This is many fewer iterations than used in the morphological snakes algorithm; however, the learned initialization φ_0 can be closer to the truth than the initial level sets given to morphological snakes. Both the initialization f_{θ_1} and update U_{θ_2} are given ResNet architectures. This allows the evolution to take advantage of learned image features. We refer to [26] for details about the architecture and loss function.

We also consider a variant of the DELSE algorithm that incorporates weak supervision. Rather than requiring the model to learn an initialization φ_0, we may initialize using a reverse buffered historical label $A(y_{it'})$, for some $t' \leq t$, as in our historically-guided U-Net and morphological snakes implementations. We only require that this model learn an update function U_θ; we remove both the network component f_{θ_1} and its associated loss.

II. Experiments

We conduct two experiments to evaluate the accuracy and computational trade-offs associated with alternative approaches to historical guidance in glacial lakes monitoring. In the first experiment, historical labels are used to support predictions across a series of timepoints in a new, moderate resolution modality. Specifically, we use imagery from Sentinel 2 from 2015 - 2016 for training, 2021 for evaluation, and 2015 - 2021 to infer overall trends. In the second experiment, these labels support prediction using a new, high resolution modality at a single future timepoint. These images are generated by MAXAR satellites and were obtained through Bing Maps in 2021.

In both experiments, we train models using the ICIMOD inventory of glacial lakes in the Himalayas [11]. These labels were collected through manual selection and correction of automatically generated hyperpixels generated on Landsat 7 imagery. The resulting labels are polygons whose vertices encode lake boundaries. A histogram of lake sizes derived from this survey is provided in Appendix Figure 9. A total of 3,624 lake boundaries are available.

Since they were curated using Landsat 7 images, these labels will not exactly match either the Sentinel 2 or Bing imagery used for training. For Sentinel 2, even after filtering to 2015 - 2016, differences in cloud cover, ice cover, and registration can lead to inconsistencies. For Bing imagery, only the most recent scenes are available, and lakes have likely changed from 2015 to 2021. Noisy labels can often lead to sensible models, which is why we train models in spite of the discrepancy [1]. However, they may lead to biased evaluations.

For this reason, we have curated additional polygon labels associated with the most recent imagery from Sentinel 2 and Bing Maps associated with randomly selected lake IDs, restricting only to those lake IDs that were not sampled by Sentinel in 2015 - 2016 or which were absent from the Bing Maps training and validation sets. Specifically, 139 and 50 lakes were randomly selected from among these Sentinel 2 and Bing Maps candidates, respectively. The most recently available lake image satisfying the same filtering criteria used in the experiments below was then labeled. Since our interest is primarily on the use of historical labels to improve predictions on more recent data sources, and since many fewer are available here compared to the ICIMOD inventory, these labels are only used for evaluation, not training.

A. Evaluation

For either set of labels, we use the following metrics to evaluate segmentation and inferred glacial lake boundary quality. Let $\hat{y} \in \{0,1\}^{H \times W}$ denote predictions for an image after thresholding associated probability assignments. Let $y \in \{0,1\}^{H \times W}$ denote ground truth. We compute the following metrics,

- Intersection-over-Union: $\frac{\sum_{ij} y_{ij}\hat{y}_{ij}}{\sum_{ij} \max(y_{ij}, \hat{y}_{ij})}$, the number of correctly labeled pixels (intersection) divided by the number of pixels belonging either to the prediction or ground truth (union).
- Precision: $\frac{\sum_{ij} y_{ij}\hat{y}_{ij}}{\sum_{ij} \hat{y}_{ij}}$, the fraction of predicted lake pixels that are correct.
- Recall: $\frac{\sum_{ij} y_{ij}\hat{y}_{ij}}{\sum_{ij} y_{ij}}$, the fraction of true lake pixels that are correctly recovered.
- Frechet Distance: Let P_V be a function that approximates the region labeled 1 in a binary mask by a polygon with V vertices. For an image u, it returns a set of vertices $p_v(u) = \left(p_v^h(u), p_v^w(u)\right) \in P_V(u)$. Then, the Frechet Distance between y and $\hat{y}$ is defined as $\inf_{\pi \in \Pi_V} \sum_{v=1}^{V} \|p_v(y) - p_{\pi_v}(\hat{y})\|^2$, where Π_V is the set of all permutations of $1 \ldots, V$.

We compute two variants of the Frechet distance. The first, Frechet (px), is based on the raw pixel coordinates of labels in the image, while the second, Frechet (m), converts distances to the physical meters between pairs of points in the image. The first approach is suited to algorithmic evaluation while the second has a clearer practical interpretation. The Frechet distance serves as a proxy for the difficulty of dragging vertices in predicted polygons to the closest location that lies on the true boundary, a task that would be necessary before releasing the final lake boundaries for downstream use. For methods that return probabilistic assignments to each pixel, we compute and report these metrics over a grid of probability thresholds from $0.05, 0.1, \ldots, 0.95$, choosing an optimal threshold using imagery in the validation set.

B. Time Course

In the first experiment, we collect time courses of Sentinel 2 imagery for each lake. Specifically, for each lake of interest, we query up to the clearest 10 images for each year from 2015 - 2021. We discard images with more than 5% cloud cover, 70% snow cover, or 20% missing values. Due to the resolution level of publicly available training imagery, only the 40% largest lakes are used in analysis. This includes all 47 that were previously designated as potentially dangerous [11]. Sample collection is more frequent in recent years —

the number of available timepoints per lake per year is given in Appendix Figure 10. Each image is cropped to a square enveloping the 10% buffered polygon associated with the 2015 lake label. If the cropped image is smaller than 500 pixels on one end, it is rescaled so that its smallest edge is 500 pixels across.

We split all imagery from 2015 - 2016 into training, validation, and test sets according to geographic basin. The basins associated with each set are chosen randomly; the correspondence is given in Appendix A. A total of 271, 47, and 85 lakes belong to these sets, respectively. The 2015 - 2016 test set is complementary to the newly labeled test set and reflects performance on images from the same period used for training. All models are trained on the same set of training images. Hyperparameters were chosen by comparing IoU on training and validation sets. Training details are given in Appendix B.

C. Modality Updating

In the second experiment, we explore the use of historical labels when generating lake boundary predictions on modality with much higher resolution than those used to generate training labels. This reflects the problem of creating labels on novel image sources when past projects have already generated labeled datasets based on previously available, potentially lower quality sources. Specifically, we download 3624 Bing Maps tiles centered around the centroids of historical lake labels. The timepoints received for each lake are those that are the most recently available as of August 2021. At the highest available resolution, some lakes do not fit into a single tile. Therefore, we adapt the zoom level of the downloaded imagery according to the 2015 lake size. This ensures that (1) large lakes are contained within the downloaded imagery and (2) smaller lakes can be viewed at high resolution.

We apply the same basin-level splits as used in the time course experiment. A total of 2128, 350, and 1146 images are available for training, validation, and testing, respectively. The discrepancy in sample sizes compared to the previous experiment reflects (1) Sentinel 2 imagery is not available for all lakes in 2015 - 2016 and (2) small lakes are not discarded in this experiment, since image resolution is higher. All hyperparameters are chosen by evaluating IoU on training and validation sets. Training details are again deferred to Appendix B.

III. Results

Table I shows the performance of the different models on Sentinel 2 and Bing Maps imagery held out test sets. The labels used for this evaluation were obtained from the ICIMOD glacial lake inventory and were created using 30m resolution Landsat 7 satellite imagery from 2015. These ground truth labels are not always accurate and/or aligned with Bing or Sentinel imagery since they were created using lower resolution images. This Sentinel 2 test set was restricted to scenes from 2015 to 2016 for better alignment with the ground truth. Since Bing Maps, does not offer historical imagery the Bing test set includes the most recent imagery for each glacial lake on the test set.

Table II shows the performance of the different models for both Bing Maps and Sentinel 2 imagery datasets. This evaluation was conducted on the held-out test set manually labeled by the authors on recent imagery.

We discuss some takeaways from these results. First, we note that, in general, no model is uniformly superior to others across all metrics. The only possible exception is the morphological snakes model applied to Bing Imagery, which is best on all metrics except recall. The model with best recall, historically-guided DELSE, has unacceptably low precision. The U-Net model generally has lower precision than other models, especially on Bing imagery. The Frechet distance is sensitive to the inclusion of false positive predictions far from the target lake prediction, explaining the poor performance of U-Net and DELSE with respect to this metric.

In general, performance on Sentinel 2 is better than for Bing. This may seem counterintuitive, because Sentinel 2 images are lower resolution. The training set for this experiment is also smaller. However, at lower resolution, lakes tend to appear more homogeneous. At higher resolution, differences in texture and color across the lake become visible — for example, deep and shallow water can be distinguished — and these more subtle variations may be difficult for models to learn. Further, for all models, performance deteriorates for the labeled imagery on recent lakes. This is expected for historically-guided models, since prior information may be less relevant when reaching further into the future. For other models, it may reflect increased difficulty of the randomly chosen recent subset.

IV. Error Evaluation

We next explore the reasons for performance differences observed in Tables I and II. In Figures 2 and 1, we visualize lake-level performance across several metrics in both experiments. Based on labels generated from recent imagery, we compute the average IoU across all models for each lake. This summarizes the overall difficulty of each lake. We then compute the $0.2, 0.4, \ldots, 0.8$ quantiles of lake performance and randomly select 5 representative lakes from each bin. Performance on these lakes is encoded as lines within each panel, and the baseline size of the lake in 2015 is encoded by the size of the points. The analogous Figures generated for the ICIMOD inventory labels and Sentinel / Bing test sets are shown in Appendix Figures 11 and 12.

For both Sentinel 2 and Bing Maps imagery evaluated with recent labels, all models achieve high performance at the 0.8 quantile (panel on the far right), except for the DELSE model with historical labels. For the remaining models, differences in average performance seems attributable to the most challenging lakes, especially those in the lowest 2 - 3 quantiles of average IoU (panels on the left). This suggests that, for lakes with the highest resolution and the clearest contours, any of these models can be applied successfully, and that the more challenging lakes in quantiles 1 - 3 are responsible for the performance differences observed in Tables I and II. For example, for both Sentinel 2 and Bing Maps imagery, the precision of the U-Net model substantially worsens in performance quantiles 1 - 3, despite having comparable precision

TABLE I
MODELS PERFORMANCE ON BING AND SENTINEL 2 IMAGERY USING HISTORICAL LABELS FROM 2015 PROVIDED BY ICIMOD

	Bing					Sentinel 2				
Model	**IoU (%)**	**Precision (%)**	**Recall (%)**	**Frechet (px)**	**Frechet (m)**	**IoU (%)**	**Precision (%)**	**Recall (%)**	**Frechet**	**Frechet (m)**
U-Net	36.5	42.3	69.0	622.04	323.46	53.0	61.5	76.6	448.81	4488
U-Net Historical	**74.7**	82.2	**90.7**	160.75	83.59	**80.2**	87.6	**91.6**	279.90	2799
DELSE	70.1	76.6	89.4	449.82	233.91	40.4	55.4	51.4	451.75	4518
DELSE Historical	53.2	**91.0**	56.5	**115.75**	**60.19**	56.7	**98.9**	57.3	83.0	830
Snake	58.9	67.1	85.8	120.38	62.60	69.3	89.1	77.6	**65.65**	**657**

TABLE II
MODELS PERFORMANCE ON BING AND SENTINEL 2 IMAGERY ON LABELED RECENT IMAGERY

	Bing					Sentinel 2				
Model	**IoU (%)**	**Precision (%)**	**Recall (%)**	**Frechet (px)**	**Frechet (m)**	**IoU (%)**	**Precision (%)**	**Recall (%)**	**Frechet (px)**	**Frechet (m)**
U-Net	42.7	48.4	72.3	629.31	327.25	62.5	68.0	84.4	408.27	4083
U-Net Historical	49.8	63.6	72.5	228.57	118.86	67.4	78.4	**84.9**	239.27	2393
DELSE	47.7	58.1	73.5	439.70	228.64	51.4	63.4	65.9	428.78	4288
DELSE Historical	37.6	**70.7**	48.7	211.32	109.89	49.9	**89.5**	54.9	130.37	1304
Snake	**52.8**	59.6	**79.8**	**159.70**	**83.04**	**70.9**	86.4	80.4	**111.09**	**1111**

to all other models in quantiles 4 - 5. Further, the historically-guided U-Net and morphological snakes models achieve high IoU even within the most challenging quantile (far left panel).

For Bing imagery, performance is generally poorest on small lakes. Lake size has relatively less influence on performance in the Sentinel 2 dataset. For Sentinel 2 imagery, the morphological snake and U-Net with labels perform comparably, though the morphological snake model typically has slightly larger variation in IoU and recall.

Tables 3 and 4 provide the predictions associated with one lake from each quantile in Figures 2 and 1, respectively. We find that the historically guided U-Net does not deteriorate as severely as other models in more challenging situations, like when images are obscured by snow or clouds (quantile 1, Figure 3) or are captured at low resolution (quantile 3, Figure 3). Morphological snake segmentations tend to have rougher boundaries than those made by U-Net and DELSE models, in spite of the smoothness regularization terms included in their objective functions. On Bing Maps imagery, it appears that the U-Net has not learned to predict lake ice regions as lake (quantiles 2 and 3, Figure 4). The historically-guided DELSE model makes only minimal changes to its initialization. This explains the high precision and low recall observed above. For this strategy to be work, it seems that a way of increasing the number of iterations used in the DELSE algorithm will need to be developed. In cases where the lake has disappeared, methods that are provided the historical label can mistakenly hallucinate the existence of a lake (quantile 1, Figure 4). That is, based on other training examples, the model has learned to always predict that the initialization must belong to the lake, despite evidence to the contrary given by the image. In fact, the morphological snakes, DELSE, and guided U-Net models grow the initialization slightly.

Only U-Net and DELSE predict non-continguous lake regions (quantiles 1 and 5, Figure 4 and quantile 4, Figure 3). This is an advantage if the goal is to discover new lakes, but may not be desirable if the goal is to update contours of a previously observed one. The DELSE predictions are much sharper on Bing Maps compared to Sentinel 2 imagery. This may reflect the fact that it includes a ResNet encoder pretrained on natural images, whose more well-defined features may be more relevant in the higher-resolution setting. It may also reflect the larger sample size in the modality updating experiment. Further examples of lake predictions corresponding to error quantiles using the ICIMOD inventory are given in Appendix Figures 13 and 14.

V. INTERACTIVE VISUALIZATION

We next use model results to analyze trends in glacial lake surface area in the HKH region. To support queries about temporal patterns, we have designed an interactive visualization system. The interface has three linked components: a selection panel, a time series view, and an imagery table. An annotated version of this interface is given Figures 5 and 6. It can also be accessed online at https://krisrs1128.shinyapps.io/glacial_lake_visualization. The selection panel is separated into three tabs. The first tab allows the user to filter down to specific basins or glacial lake IDs of interest. The lake selection options are updated conditionally on the basins chosen. The second tab allows the user to filter to selected lake IDs. The third tab allows the user to choose a trend of interest and the lake ID selection options also updates accordingly. The range of area shown in the time series can be adjusted by moving the slider bar at the top of the selection panel. All the selections can be reset by pressing the "Reset" button at the bottom of the selection panel.

Below the selection panel, the time series view shows inferred lake areas over time from the predictions of the historically-guided U-Net, though in principle, any source of label predictions could be substituted. When the user brushes lakes within this plot, the color of the brushed point is changed to red and the associated imagery is displayed in the table below. Each row of this table corresponds to one lake; columns give timepoints. Hovering over an image shows the specific date at which the scene was sensed. The interface is designed to support exploration of lake images with properties of interest, including geographical location and trends in inferred size.

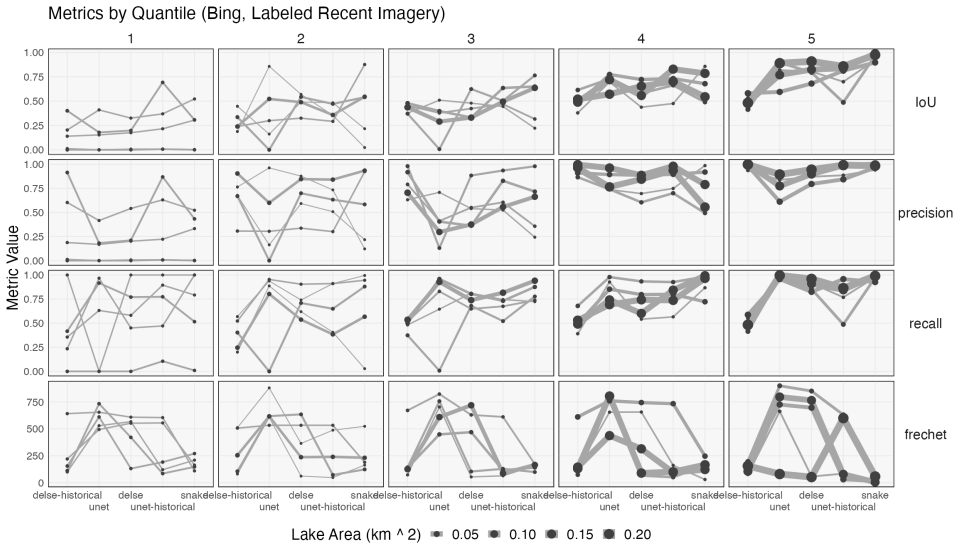

Fig. 1. A comparison of lake-wise performance across models when evaluating with newly labeled, high-resolution imagery. Each column corresponds to one quantile of lake performance (lowest to highest performance from left to right), and each row is a performance metric. The five lines in each panel correspond to five randomly selected lakes within that quantile. Models are sorted from those with lowest average IoU to those with highest average IoU.

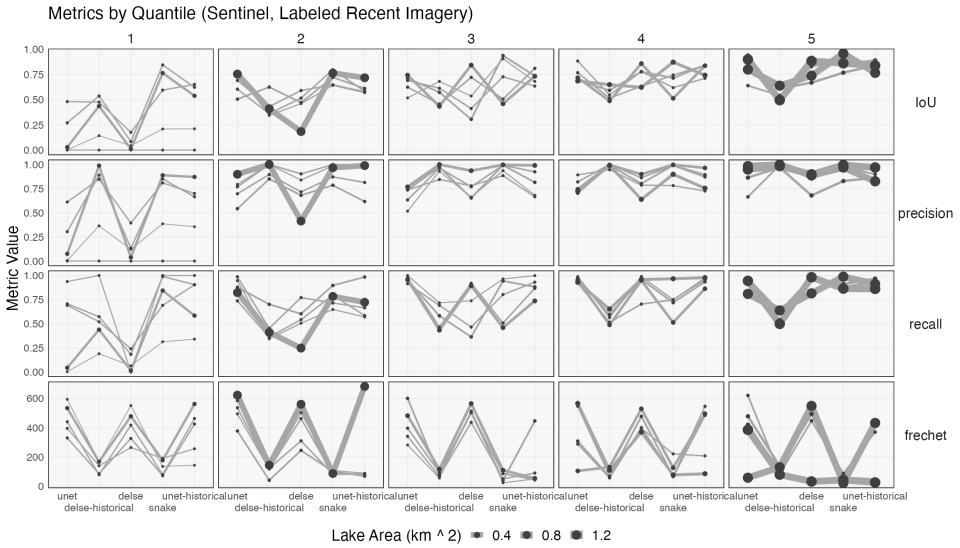

Fig. 2. The analog of Figure 1 generated for newly labeled Sentinel 2 imagery. Models are again sorted from those with lowest to highest average IoU, though note that the order has changed.

VI. Trend Analysis

To summarize trends in predicted lake areas, we have fitted a collection of linear regression models between sample collection date and predicted lake areas from the historically-guided U-Net. Specifically, for each lake with at least four observations, we fit the model

$$\log\left(\text{Area}_i\right) = \beta_{\text{intercept}} + \beta_{\text{slope}}\text{Date}_i,$$

where Date_i counts the number of days since 01/01/2015.

The estimated $\hat{\beta}_{\text{slope}}$ terms describe the linear component of any trends that may exist for each lake. Figure 7 is a volcano plot of the estimated effects, displaying standardized slopes against their associated p-value. Estimates made using a wider temporal range and with more samples will have lower standard errors and more significant p-values. Points in the upper left and right corners correspond to glacial lakes where a large, significant effect has been detected. The fact that more negative $\hat{\beta}_{\text{slope}}$ estimates are observed suggests that more glacial lakes have decreasing than increasing areas. However, we note that the point close to $(0, 0)$ contains the majority of glacial lakes – most lakes do not change in any detectable way over the time span of the experiment.

Two of the lakes with significant increasing trends are shown in Figure 8. The lower boundary of glacial lake `GL086379E28392N` has clearly expanded while lake `GL087401E28768N` has grown in all directions.

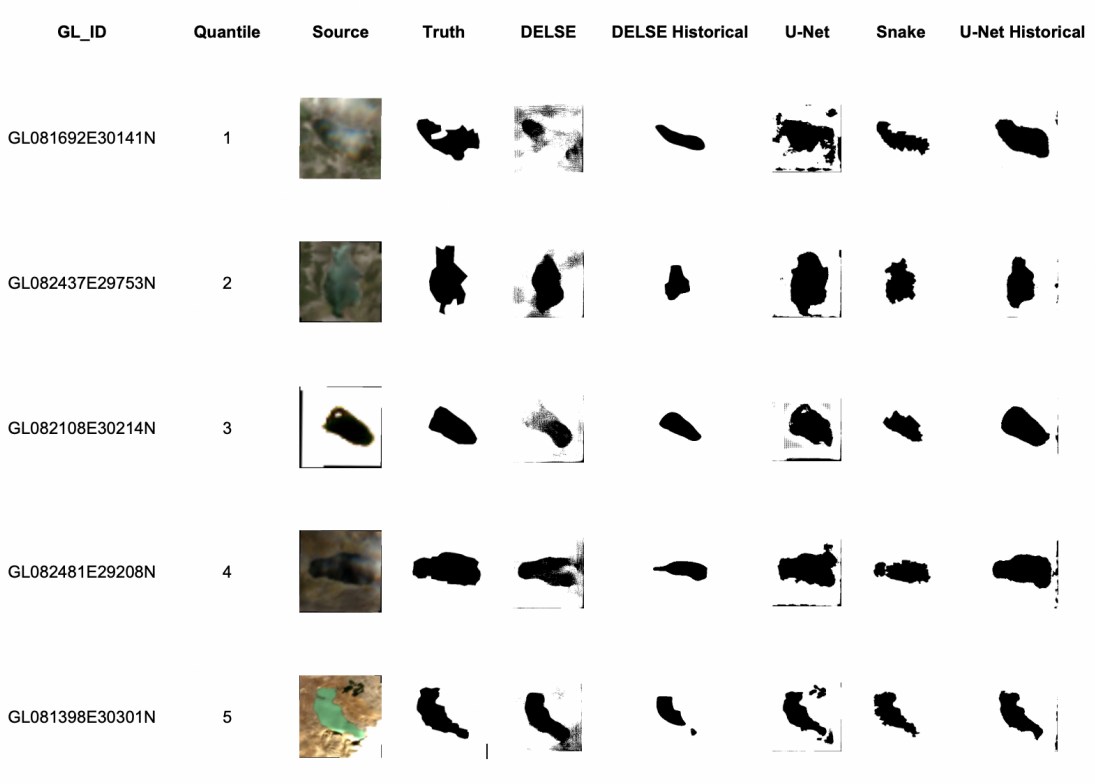

Fig. 3. A comparison of errors made across models on Sentinel 2 imagery. Models are arranged along columns. Each row provides a randomly selected lake whose average error lie between the first (highest error, top row) through fifth (lowest error, bottom row) quantiles. In these images, regions with clouds (row 1) or at low resolution (row 3) tend to lead to elevated error in models without access to historical labels. The GL_ID column provides glacial lake IDs as defined in ICIMOD's 2015 inventory [11].

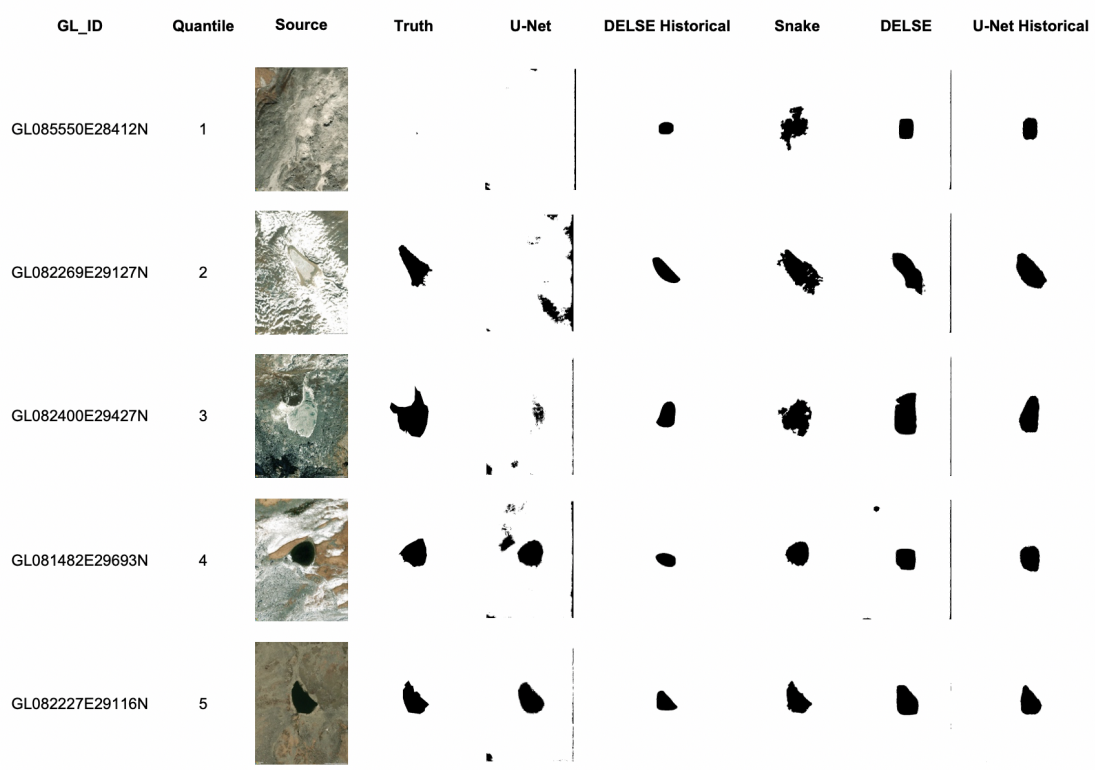

Fig. 4. A comparison of errors made across models on high-resolution Bing imagery. Models and average error rates are arranged as in Table 3. The morphological snakes model achieves the highest average IoU on this dataset.

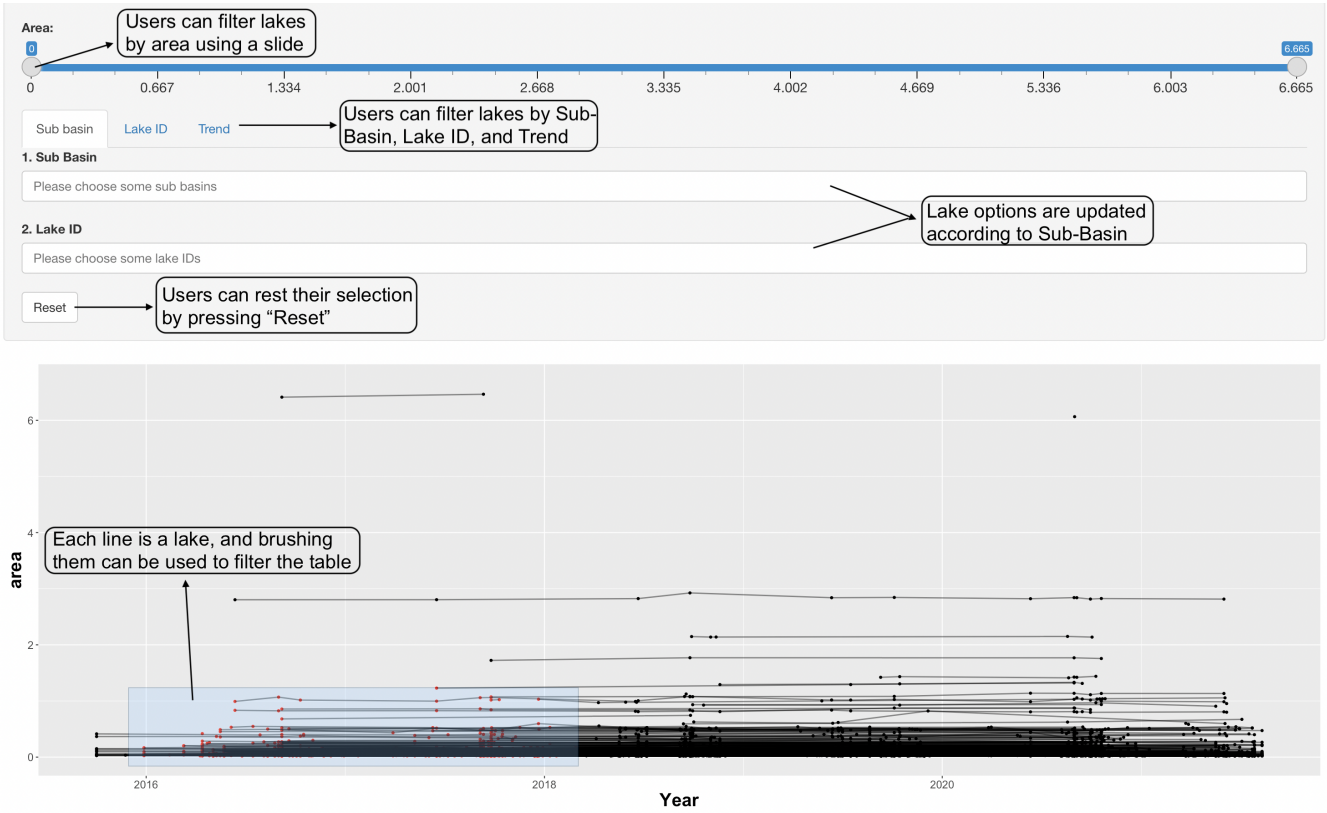

Fig. 5. A screenshot from the selection panels of the trend analysis visualization application. Users can select lakes for inspection using either static properties or model-predicted lake areas. The selected lakes populate the table shown in Figure 6.

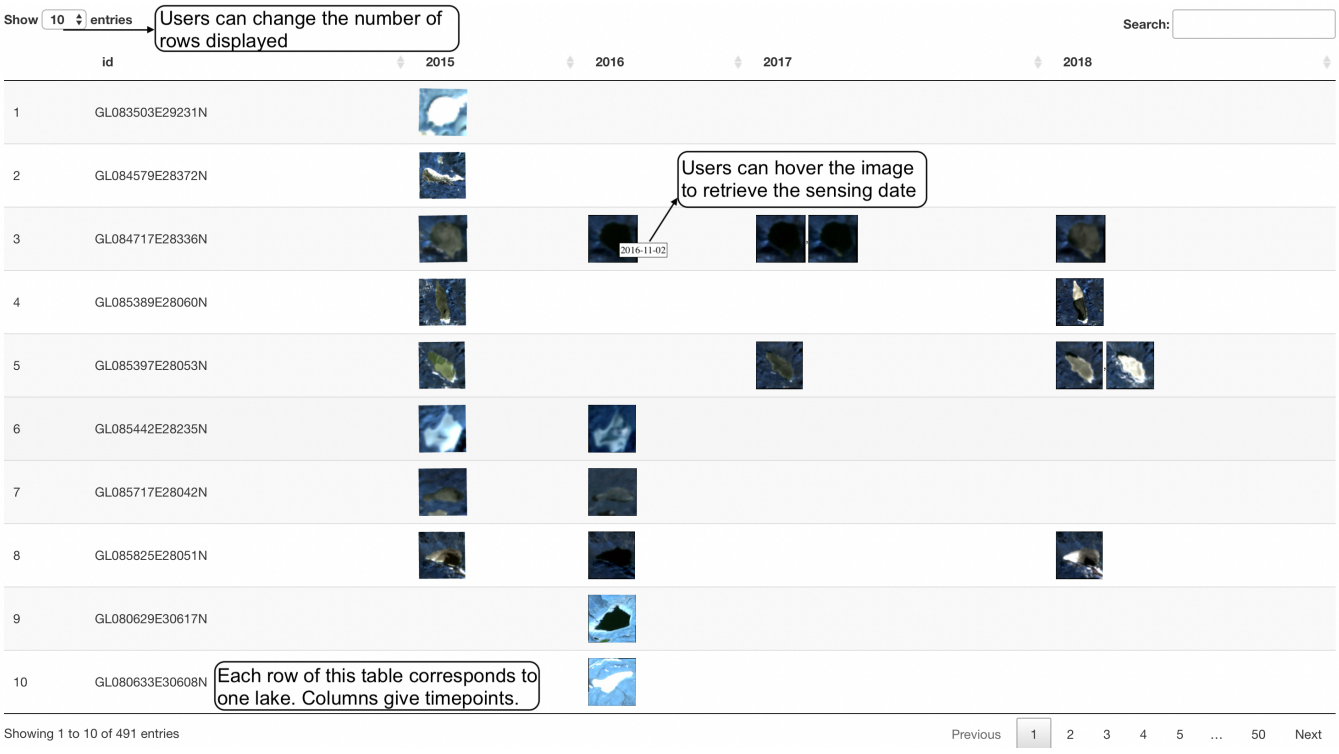

Fig. 6. A screenshot from the lake time series visualization from the trend analysis visualization application. Given a highlighted set of lakes, the table updates to show the corresponding time series of imagery. The application provides a structured approach to exploring model-predicted lake areas.

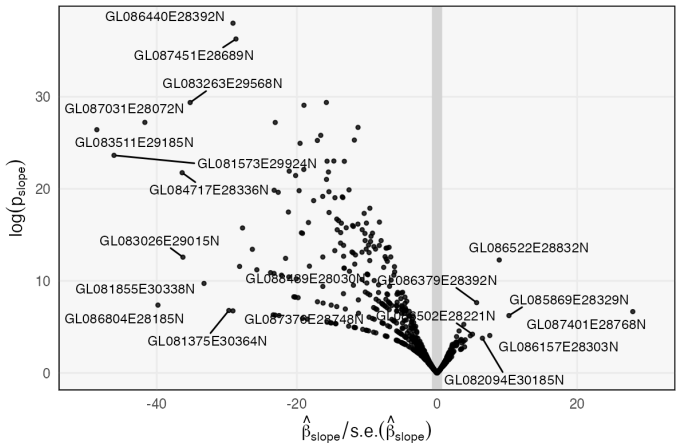

Fig. 7. A volcano plot of estimated glacial lake area changes. Lakes at the bottom of the "V" do not have detectable trends in glacial lake area. Those on the upper right (left) have significant increasing (decreasing) areas. Two example glacial lakes labeled here are shown in Figure 8.

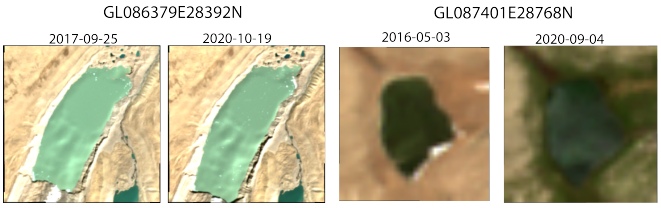

Fig. 8. Example Sentinel-2 imagery of growing glacial lakes identified using Figure 7 and the interactive visualization tool. Only two timepoints are displayed. The image for lake `GL086379E28392N` appears higher resolution because the original lake area is larger.

VII. Discussion

We have studied methods for incorporating historical labels to guide glacial lake segmentation on more recently sampled imagery. Our work provides a template for earth observation tasks that could benefit from transferring labels from one modality to another. We have applied both state-of-the-art level set evolution methods and simple preprocessing strategies. For Sentinel 2 imagery, we have found that a concatenation of partially obscured historical labels as an extra channel to U-Net performs well, mainly by reducing the number of false positive assignments to shadow or snow, since these potential distractors tend to be far from historical lake labels. For Bing Maps imagery, we found that initializing a morphological snake at a shrunken version of the historical label was effective. However, both approaches could be misled to produce false positives segmentations when the true lakes are no longer enclosed in the initialization.

We analyzed predictions from the historically-guided U-Net model on Sentinel 2, identifying several glacial lakes whose areas have noticeably increased since the 2015, the last time the glacial lake inventory was formally curated. Though most lakes have remained relatively stable over the time span of our data, we observed more lakes with decreasing, rather than increasing, trends. To more easily analyze lake area trends, we developed an interface to support visual queries. We believe these methods will support more frequent inventory and analysis of glacial lake areas in the HKH, providing source material for GLOF risk assessment.

In practice, not all lakes of interest have been captured by historical surveys. Future work will concentrate on a closer coupling between discovery of new lakes and boundary updating for previously identified ones. For example, a model for detecting lakes can be coupled with an interface for supplying and automatically refining weak, partial labels, similar to those used to guide predictions in this study.

Further, we observed several limitations of existing methods, suggesting avenues for further study. All models except for the morphological snake model struggled with high-resolution imagery, where a single lake might include several subtly distinguishable textures. The highly-parameterized DELSE model failed to outperform a simple modification of U-Net, perhaps due to the limited sample size of the datasets studied.

Code for both modeling and trend analysis is available in these github repositories [Models, Visualizations, Experiments]. With the exception of proprietary Bing Maps imagery used in the modality updating experiment, all data, labels, and predictions associated with this study are available. Links are given in Appendix C.

Acknowledgment

This research was performed using the compute resources and assistance of the UW-Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation. The authors thank Microsoft and the AI for Earth initiative for their support with Microsoft Azure resources and access to Bing Maps imagery.

References

[1] Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771, 2021.

[2] Luis Álvarez, Luis Baumela, Pedro Henríquez, and Pablo Márquez-Neila. Morphological snakes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2197–2202. IEEE, 2010.

[3] LM Andreassen, F Paul, A Kääb, and JE Hausberg. Landsat-derived glacier inventory for jotunheimen, norway, and deduced glacier changes since the 1930s. *The Cryosphere*, 2(2):131–145, 2008.

[4] Jonathan L Carrivick and Fiona S Tweed. A global assessment of the societal impacts of glacier outburst floods. *Global and Planetary Change*, 144:1–16, 2016.

[5] Fang Chen. Comparing methods for segmenting supraglacial lakes and surface features in the mount everest region of the himalayas using chinese gaofen-3 sar images. *Remote Sensing*, 13(13):2429, 2021.

[6] Stephan Harrison, Jeffrey S Kargel, Christian Huggel, John Reynolds, Dan H Shugar, Richard A Betts, Adam Emmer, Neil Glasser, Umesh K Haritashya, Jan Klimeš,

et al. Climate change and the global pattern of moraine-dammed glacial lake outburst floods. *The Cryosphere*, 12(4):1195–1209, 2018.

[7] Yi He, Sheng Yao, Wang Yang, Haowen Yan, Lifeng Zhang, Zhiqing Wen, Yali Zhang, and Tao Liu. An extraction method for glacial lakes based on landsat-8 imagery using an improved u-net network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:6544–6558, 2021.

[8] Jack D Ives, Rajendra B Shrestha, Pradeep K Mool, et al. *Formation of glacial lakes in the Hindu Kush-Himalayas and GLOF risk assessment*. ICIMOD Kathmandu, 2010.

[9] Sanjay K Jain, Anil K Lohani, RD Singh, Anju Chaudhary, and LN Thakural. Glacial lakes and glacial lake outburst flood in a himalayan basin using remote sensing and gis. *Natural hazards*, 62(3):887–899, 2012.

[10] Junli Li and Yongwei Sheng. An automated scheme for glacial lake dynamics mapping using landsat imagery and digital elevation models: A case study in the himalayas. *International Journal of Remote Sensing*, 33(16):5194–5213, 2012.

[11] Sudan Bikash Maharjan, PK Mool, W Lizong, G Xiao, F Shrestha, RB Shrestha, NR Khanal, SR Bajracharya, S Joshi, S Shai, et al. *The Status of Glacial Lakes in the Hindu Kush Himalaya-ICIMOD Research Report 2018/1*. International Centre for Integrated Mountain Development (ICIMOD), 2018.

[12] Pradeep K Mool, Dorji Wangda, Samjwal R Bajracharya, Karma Kunzang, Deo R Gurung, Sharad P Joshi, et al. Inventory of glaciers, glacial lakes and glacial lake outburst floods. monitoring and early warning systems in the hindu kush-himalayan region: Bhutan. *Inventory of glaciers, glacial lakes and glacial lake outburst floods. Monitoring and early warning systems in the Hindu Kush-Himalayan Region: Bhutan.*, 2001.

[13] Fan Mou, Danyang Wang, Jiaxi Liu, Zezhong Zheng, Liming Jiang, Guoqing Zhou, and Fangrong Zhou. Change of glacial lake in karakoram range. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 2539–2542. IEEE, 2020.

[14] Fakhra Muneeb, Siddique Ullah Baig, Junaid Aziz Khan, and Muhammad Fahim Khokhar. Inventory and glof susceptibility of glacial lakes in hunza river basin, western karakorum. *Remote Sensing*, 13(9):1794, 2021.

[15] Yong Nie, Yongwei Sheng, Qiao Liu, Linshan Liu, Shiyin Liu, Yili Zhang, and Chunqiao Song. A regional-scale assessment of himalayan glacial lake changes using satellite observations from 1990 to 2015. *Remote Sensing of Environment*, 189:1–13, 2017.

[16] Sven Nussbaum and Gunter Menz. ecognition image analysis software. In *Object-based image analysis and treaty verification*, pages 29–39. Springer, 2008.

[17] Stanley Osher and Nikos Paragios. *Geometric level set methods in imaging, vision, and graphics*. Springer Science & Business Media, 2003.

[18] Maxim A Petrov, Timur Y Sabitov, Irina G Tomashevskaya, Gleb E Glazirin, Sergey S Chernomorets, Elena A Savernyuk, Olga V Tutubalina, Dmitriy A Petrakov, Leonid S Sokolov, Mikhail D Dokukin, et al. Glacial lake inventory and lake outburst potential in uzbekistan. *Science of the Total Environment*, 592:228–242, 2017.

[19] Nida Qayyum, Sajid Ghuffar, Hafiz Mughees Ahmad, Adeel Yousaf, and Imran Shahid. Glacial lakes mapping using multi satellite planetscope imagery and deep learning. *ISPRS International Journal of Geo-Information*, 9(10):560, 2020.

[20] Shaun D Richardson and John M Reynolds. An overview of glacial hazards in the himalayas. *Quaternary International*, 65:31–47, 2000.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[22] T Strozzi, A Wiesmann, A Kääb, S Joshi, and P Mool. Glacial lake mapping with very high resolution satellite sar data. *Natural Hazards and Earth System Sciences*, 12(8):2487–2498, 2012.

[23] Jinro Ukita, Chiyuki Narama, Takeo Tadono, Tsutomu Yamanokuchi, Nobuhiro Tomiyama, Sachi Kawamoto, Chika Abe, Tsuyoshi Uda, Hironori Yabuki, Koji Fujita, et al. Glacial lake inventory of bhutan using alos data: methods and preliminary results. *Annals of Glaciology*, 52(58):65–71, 2011.

[24] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[25] Xin Wang, Xiaoyu Guo, Chengde Yang, Qionghuan Liu, Junfeng Wei, Yong Zhang, Shiyin Liu, Yanlin Zhang, Zongli Jiang, and Zhiguang Tang. Glacial lake inventory of high-mountain asia in 1990 and 2018 derived from landsat images. *Earth System Science Data*, 12(3):2169–2182, 2020.

[26] Zian Wang, David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Object instance annotation with deep extreme level set evolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7500–7508, 2019.

[27] Renzhe Wu, Guoxiang Liu, Rui Zhang, Xiaowen Wang, Yong Li, Bo Zhang, Jialun Cai, and Wei Xiang. A deep learning method for mapping glacial lakes from the combined use of synthetic-aperture radar and optical satellite images. *Remote Sensing*, 12(24):4020, 2020.

[28] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

APPENDIX A
DATA DETAILS

The mapping between train, validation, and test basins is given below,

- Train: Arun, Bheri, Budhi Gandaki, Dudh Koshi, Humla, Indrawati, Kali, Kali Gandaki
- Validation: Karnali, Kawari, Likhu, Marsyangdi, Mugu, Seti
- Test: Sun Koshi, Tama Koshi, Tamor, Tila, Trishuli, West Seti

APPENDIX B
TRAINING DETAILS

Below, we provide model training details for the time course experiment.

- For both U-Net and DELSE models, one epoch consists of looping over every image in the training set, in random order. For each image, a random 400×400 patch is selected.
- The U-Net models were trained for 200 epochs with batch size of 8.
- For the U-Net models, a Stochastic Gradient Descent optimizer with learning rate 5e-4 was used.
- The DELSE models were trained for 200 epochs with a batch size of 2 and 3 refinement iterations per batch.
- A weighted cross-entropy loss was used for training U-Net.
- For both U-Net and DELSE models, gradient clipping was used with a maximum gradient norm of 10.
- Each head of the DELSE model was trained for 10000 iterations before joint training.
- For DELSE models, a Stochastic Gradient Descent optimizer with learning rate 1e-4 was used.

Below, we provide training details for the modality updating experiment.

- For both U-Net and DELSE models, one epoch consists of looping over every image in the training set, in random order. For each image, a random 400×400 patch is selected.
- The U-Net models were trained for 60 epochs with batch size of 8.
- For the U-Net models, a Stochastic Gradient Descent optimizer with learning rate 5e-4 was used.
- The DELSE models were trained for 60 epochs with a batch size of 2 and 3 refinement iterations per batch.
- A weighted cross-entropy loss was used for training U-Net.
- For both U-Net and DELSE models, gradient clipping was used with a maximum gradient norm of 10.
- Each head of the DELSE model was trained for 6000 iterations before joint training.
- For DELSE models, a Stochastic Gradient Descent optimizer with learning rate 3e-4 was used.

APPENDIX C
DATA DOCUMENTATION

Download scripts

- All Sentinel 2 imagery were downloaded through the Planetary Computer using this script and the helper functions here.

Vector Label Data

- ICIMOD Inventory: Vector labels curated by the 2015 inventory are available in the ICIMOD regional database at this link.
- Curated Evaluation Labels: Labels on recent Sentinel 2 and Bing imagery used for evaluation in this study are available at this link.

Images and Rasterized Labels

- Sentinel: Sentinel 2 imagery with ICIMOD inventory labels are given in `sentinel.tar.gz`. The subfolder `splits` contains the resized, transformed images associated with the training, validation, and testing splits. Each split is contained in its own folder. Each image can be associated with a rasterized label in the `labels` subfolders of the raw and split data.
- Sentinel 2 imagery and labels from the review conducted by the authors are available at this link.
- Imagery from Bing Maps are proprietary and cannot be shared publicly.

Predictions

- Metrics for all models, lakes, and train / validation / test subsets at all probability thresholds are given in this folder.
- Predictions for all models and datasets are saved as both raster and vector data in zipped archives this folder. The name of the archive specifies the name of the dataset and the model.
- Predictions converted to black and white images, which were used in the error analysis above, are stored in this archive.

APPENDIX D
REPRODUCIBILITY

A Docker image with all necessary software preinstalled is available on dockerhub. It includes the packages listed in this install script.

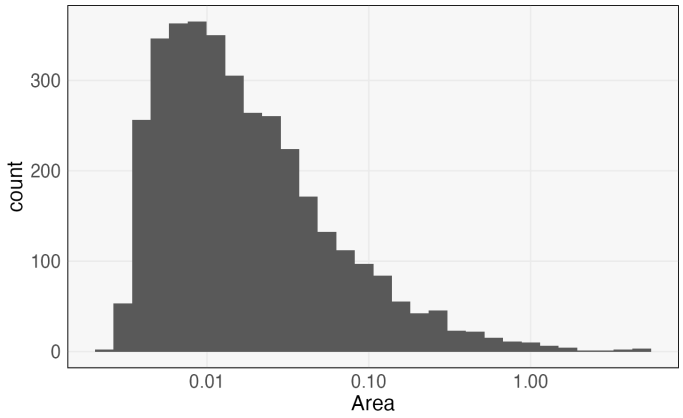

Fig. 9. Lake areas derived from the 2015 survey [11]. Note that the x-axis is on a log scale. Only the top 40% of lakes are kept, corresponding to an area cutoff of 0.019.

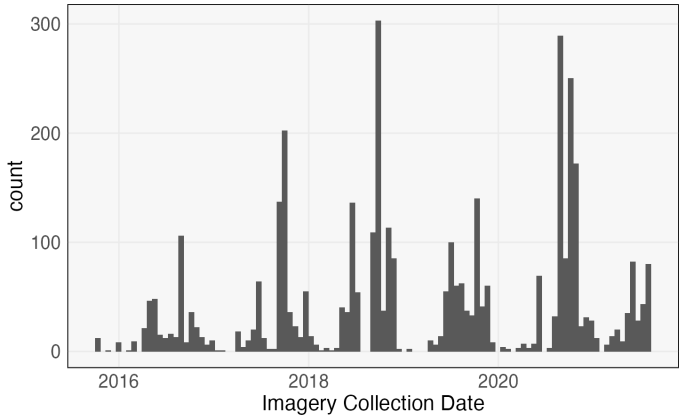

Fig. 10. Sentinel 2 imagery collection date. More imagery is available in recent years. Since labels were created in a 2015 study, only images from 2015 - 2016 were used for training and evaluation.

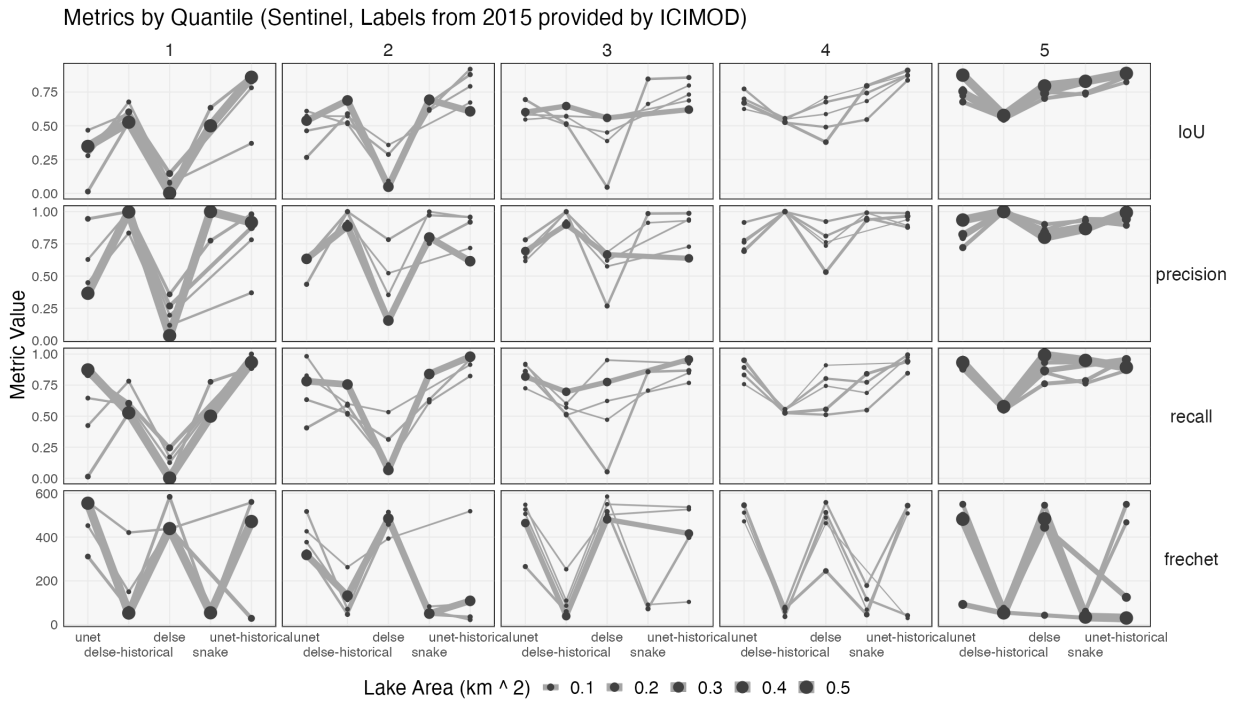

Fig. 11. The analog of Figure 2 for the error with respect to 2015 ICIMOD labels for lakes within the Sentinel test set.

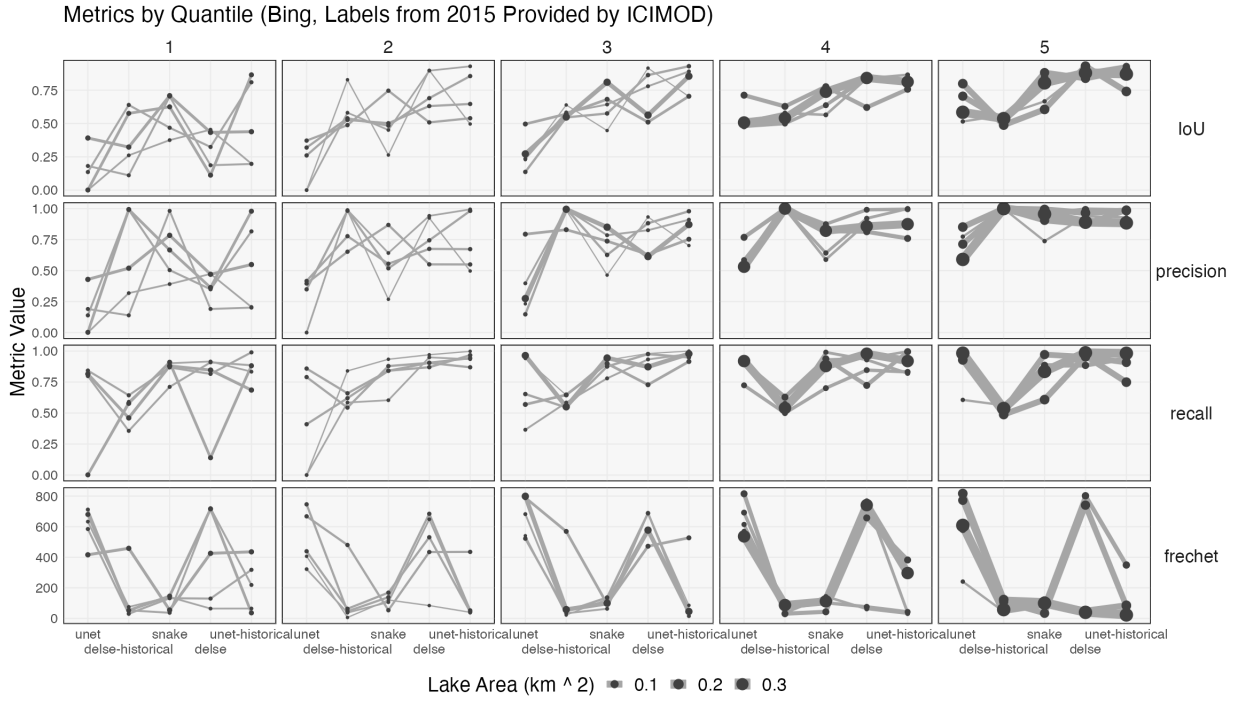

Fig. 12. The analog of Figure 1 for the error with respect to 2015 ICIMOD labels for lakes within the Bing test set. These 2015 labels are more plentiful, but not as accurate on the recently collected Bing imagery.

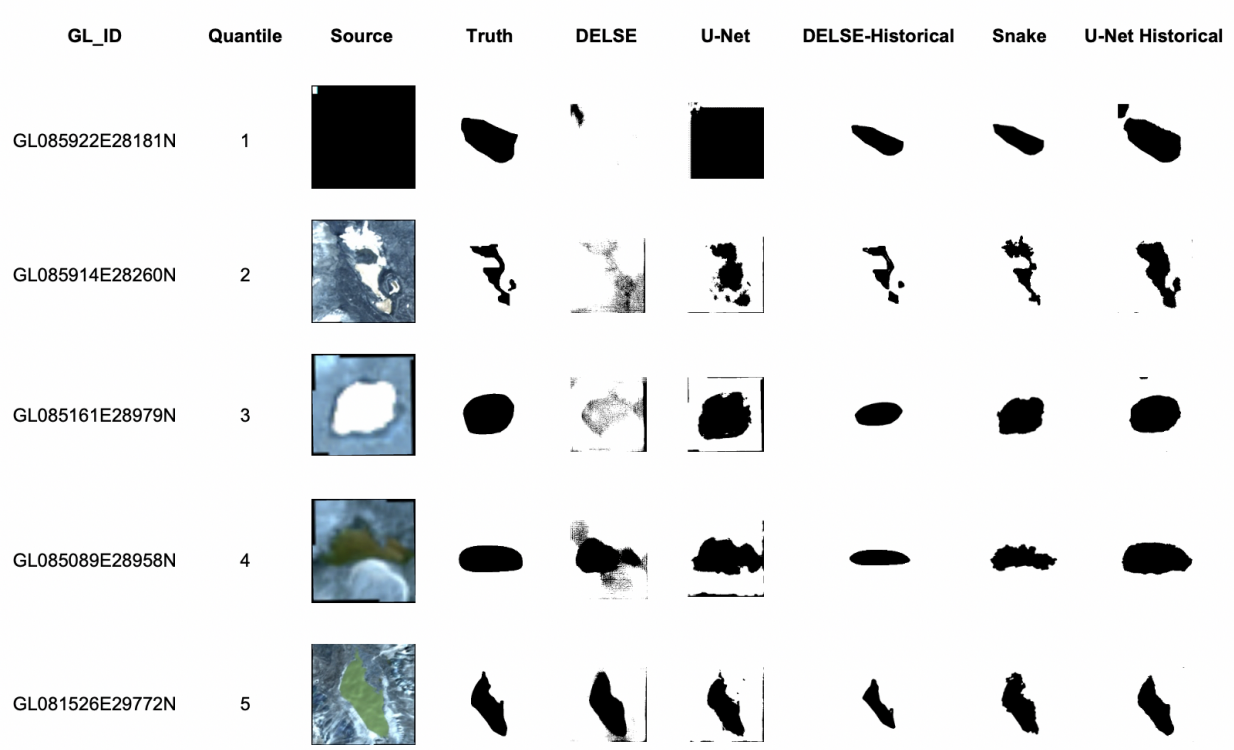

Fig. 13. Example lake predictions taken from the quantiles shown in Appendix Figure 11. The ground truth shown here are those from the 2015 ICIMOD survey. The lake from the highest error quantile is masked by a no-data region of a Sentinel 2 tile.

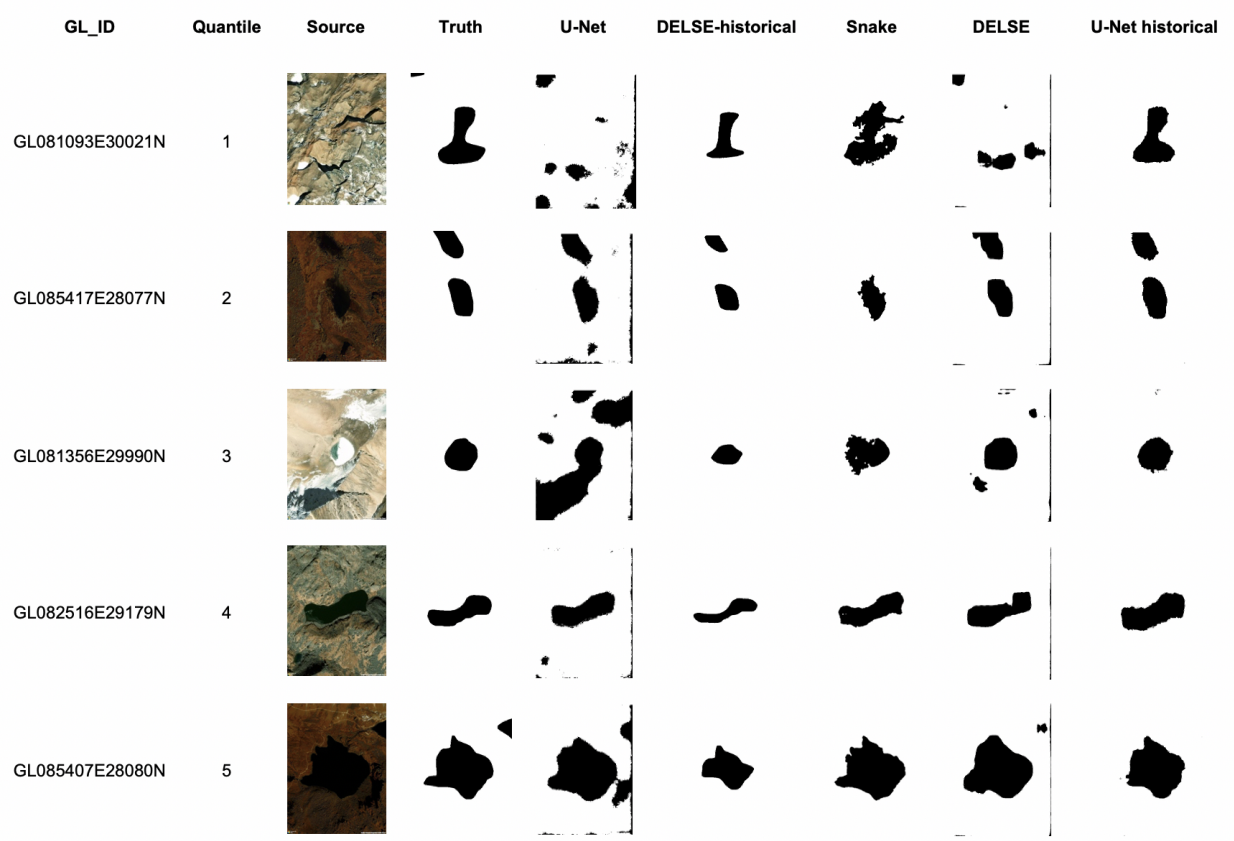

Fig. 14. Example lake predictions taken from the quantiles shown in Appendix Figure 12. These are the labels that were used to train the models in the Modality Updating experiment in section II. Note that labels sometimes no longer match the associated image; for example, in the first row, the lake has disappeared. This is because the labels from this inventory were prepared using Landsat 7.