Responses to Reviewer #1

Main Comment

The authors state that they used "two cloud-free WorldView-2 orthorectified images with a spatial resolution of 0.4 meters, acquired on November 3, 2012, and September 26, 2021." However, they later clarify that only the panchromatic (PAN) band is available at this resolution, while they appear to use the color bands instead. This is unclear—did they use pansharpening? Please clarify which bands were actually used, at what resolution, and whether pansharpening was applied.

Response: We appreciate this critical observation. Our study obtained the 8-band multispectral WorldView-2 imagery, which originally had a spatial resolution of 1.64 m, along with the 0.41 m panchromatic image. We applied a pansharpening process to the multispectral bands to enhance spatial detail, resulting in an effective spatial resolution of 0.4 m for all used spectral bands. This pansharpened dataset was used for all feature extraction, including vegetation indices and texture features. We revised the manuscript as follows (L116-119):

> "To enhance spatial detail, all multispectral bands were pansharpened using the corresponding high-resolution panchromatic band, yielding a uniform spatial resolution of 0.4 meters across all datasets used for feature extraction. The pansharpened multispectral imagery was the basis for deriving vegetation indices and texture features."

The origin of the training data is not clearly explained. The authors write: "Ground truth data in the study area were labeled using a pixel-based approach and categorized into four classes: bare land, forest, krummholz, and shadow (Fig. 3)." Does this mean an operator manually classified these images? If both images were already classified, what is the purpose of the complex processing workflow? Were both images used for training? If only one image was used for training, why would we expect the same classification accuracy to transfer to the second image, especially given possible environmental and seasonal differences?

Response: Yes, the ground truth was manually labeled and validated by trained operators and domain experts using a pixel-based approach (Fig. 3) in the 2021 WorldView-2 orthorectified images. The 2021 images were used for model training and validation, while the 2012 images were used for model evaluation. The classification results of 2012 and 2021 images were compared for temporal change analysis.

The purpose of the complex processing workflow was hoping to establish a universal model that can provide reliable alpine treeline ecotone classification. Only 2021 images were used for model training. We did not expect to see the same classification accuracy. Instead, we evaluated the 2012 image classification results carefully and the results were used to study the temporal and spatial changes.

The environmental and seasonal differences were minimized by using the autumn images when vegetation in the area had entered dormancy with less phenological presentation such as flowering or leaf flushing.

Additionally, we applied histogram matching during preprocessing to reduce radiometric and color inconsistencies caused by differences in lighting and atmospheric conditions. We revised the manuscript as follows (L120-124):

"Two orthorectified, cloud-free WorldView-2 images acquired on November 3, 2012, and September 26, 2021, were obtained from RiChi Technology Co., Ltd. (New Taipei City, Taiwan). Both images were captured in the autumn season when vegetation had entered dormancy, minimizing the influence of phenological variability such as flowering. Histogram matching was applied to ensure radiometric consistency across the two images. In addition, GPS devices were used to record field survey points, which were subsequently used to verify treeline positions and assist in manual ground truth labeling."

Regarding training, the authors mention using 512x512 patches and then splitting the dataset. Is the train/test split done at the patch level or at the pixel level (within patches)? This distinction is important, as pixel-level splits can introduce data leakage, especially in spatially autocorrelated datasets.

Response: Thank you for pointing this out. The dataset was split at the patch level, not the pixel level. We revised the manuscript as follows (L201-204):

"Each image (5380 × 4671 pixels) was segmented into 110 non-overlapping patches of 512 × 512 pixels. The dataset split was performed at the patch level, not the pixel level, to avoid spatial autocorrelation and data leakage. Specifically, 80% of the patches were randomly selected for training and validation (with a 75/25 split), and the remaining 20% were used as an independent test set. In total, 66 patches were used for training, 22 for validation, and 22 for testing."

The use of Random Forest (RF) for variable importance analysis is questionable. This approach is valid only if variables are independent, which is clearly not the case here. Additionally, is it worth performing this complex selection to save 20% of variables? Reducing from 77 to 61 features may not justify the effort, especially if interpretability or performance gain is marginal. As such, the entire discussion about variable importance remains inconclusive.

Response: Thank you for your comment. We acknowledge that Random Forest (RF) variable importance measures can be biased when input features are correlated. In our dataset, some of the 77 features—such as vegetation indices and texture metrics—are derived from overlapping spectral bands and are therefore not entirely independent. Nevertheless, RF remains a widely used method for feature selection in high-dimensional remote sensing and ecological data, and its robustness has been demonstrated even in the presence of correlated variables (Cutler et al., 2007; Belgiu & Drăguț, 2016). In our study, RF was employed primarily to rank features and facilitate a conservative feature selection process, ultimately retaining the top 61 out of 77 features. This reduction resulted in a 14.3% decrease in training time and a slight improvement in overall accuracy (OA increased from 0.838 to 0.842). Although the gains were modest, this optimization was valuable considering the computational demands of the U-Net model.

Reference:

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J.: Random forests for classification in ecology. Ecology, 88(11), 2783-2792, https://doi.org/10.1890/07-0539.1, 2007

Belgiu, M., and Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens., 114, 24-31, https://doi.org/10.1016/j.isprsjprs.2016.01.011, 2016.

Finally, the reported 14-meter height increase lacks context. The sentence "Forest area and highest point height difference from 2012 to 2021" is vague. Does this mean the authors extracted the maximum elevation value among all forest pixels? What was done to ensure robustness against outliers or noise? Also, scientific results are typically reported with associated uncertainties, which are missing here—or, if included, were not clear to me.

Response: Thanks for your comment. The 14-meter elevation gain was calculated based on the difference between the highest elevations of the forest area in 2021 and 2012. To elaborate on the finding, we have performed another analysis by investigating the elevation percentage of forest cover in 2021 and 2012. Based on the results, we revised the manuscript as follows (L291-296):

> “Based on the 95th percentile of DEM elevation values of all pixels classified as forest (Fig. 9), the treeline showed an upward shift of 32.00 meters between 2012 and 2021. The 95% confidence interval (± 4.00 meters) was estimated using a bootstrap resampling method (5,000 iterations). Differences in area changes across various elevation ranges are detailed in Table 8, with the most significant changes occurring in the 3,500- to 3,600-m range. The most stable area was observed in the range of 3,700 to 3,800 m.”

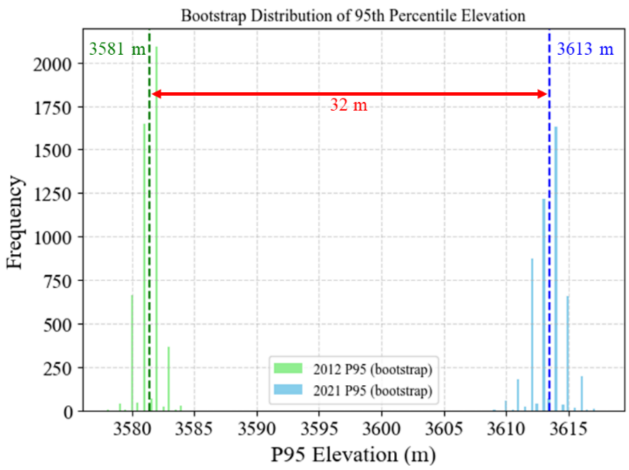

Figure 9. Bootstrap distribution of the 95th percentile elevation of forest cover for 2012 and 2021. The histogram shows the frequency of estimated 95th percentile elevations (P95) based on resampling. Green bars represent 2012 estimates, while blue bars represent 2021. The dashed vertical lines indicate the mean P95 value for each year.

Lastly, if the only interest was in changes to forest cover, why not classify the change directly instead of classifying each image independently?

Response: Thank you for the comment. We manually labeled each image. However, only the 2021 image was used for model training. The 2012 image was classified using the trained model, and the labeling for 2012 was done afterward for accuracy validation, not training. Direct change classification would require labels from both years or a different type of model. Our goal was to evaluate whether a model trained on recent data could still perform well on earlier imagery. Since both images were taken in autumn and we applied histogram matching, seasonal and lighting differences were minimized.

Minor Comments

"Taiwan has the highest density of high mountains globally, with over 200 peaks exceeding 3,000 meters in elevation."

→ This sounds too subjective. The result depends on the threshold chosen. I recommend rewriting as:

"Taiwan is one of the regions with the highest density of high mountains, with over 200 peaks exceeding 3,000 meters in elevation."

Response: Thank you for the comment. We revised the manuscript as follows (L13-14):

> "Taiwan is characterized by high mountains density, with over 200 peaks exceeding 3,000 meters in elevation."

The introduction goes beyond the immediate scope of the study. However, I appreciate that the authors took the time to place their work in a broader context.

"At the same time, the productivity of alpine treeline vegetation increased, enhancing the ability to sequester atmospheric CO_2 and mitigating the effects of climate change (Rumpf et al., 2022)"

→ If this is true, however it's also be stated that the global effect is likely minor. The sentence could be more balanced.

Response: Thank you for the comment. We revised the manuscript as follows (L324-326):

> "In contrast, studies in the European Alps have noted significant reductions in snow cover and increased alpine vegetation productivity, potentially enhancing local carbon sequestration, although with a limited global impact (Rumpf et al., 2022)"

Reference:

Rumpf, S. B., Gravey, M., Brönnimann, O., Luoto, M., Cianfrani, C., Mariethoz, G., and Guisan, A.: From white to green: Snow cover loss and increased vegetation productivity in the European Alps. Science, 376(6597), 1119-1122, https://doi.org/10.1126/science.abn6697, 2022.

Responses to Reviewer #2

Dear authors,

It was a pleasure to read your well-written manuscript. You used two machine-learning (ML) methods to classify four land-cover/ image section classes (forest, krummholz, bare ground and shadow) in a montane-alpine transition in Taiwan, repeating the classification for images from two years to detect changes through time.

The machine-learning methods are very nicely explained. However, I am missing information about how you obtained your ”ground truth”. This is actually not collected on the ground or by manual/visual labelling, if I interpreted the flowchart correctly, but by automated (?) pixel-based classification (not further specified…). So it sounds like you use one classification method to validate another, which seems strange and a bit circular. Why did you not just use the pixel-based classification for both images, if that worked so well that you could use it as “ground truth”? Why the step of developing the ML methods?

Response: We appreciate the comments and would like to clarify the misunderstanding. The ground truth labels in our study were confirmed through manual image digitization and expert ground surveys. These manually annotated labels can be used as independent ground truth for training, validating, and testing in machine learning models. Therefore, we revised the manuscript as follows (L193-200):

> “Ground truth data for the study area were manually labeled using a pixel-based approach and categorized into four classes: (1) Bare land, referring to areas of exposed soil, rock surfaces, or sparsely vegetated ground; (2) Forest, defined as regions with dense, continuous tree canopy cover; (3) Krummholz, representing stunted, shrub-like trees typically found at high elevations near the treeline and shaped by wind or snow pressure (Liao et al., 2023a); and (4) Shadow, representing regions with low reflectance caused by topographic shading or solar angle effects. The class definitions were established based on visual inspection and field knowledge of the study area (Fig. 3). The labeling process was independent and performed by visually interpreting the pansharpened RGB composite imagery, referencing known terrain characteristics, and assisted by field-collected GPS survey points.”

Pixel-based labels are used solely for model training and evaluation and not for generating classification results. Due to the complexity of the alpine treeline ecotone, spectral signals are often affected by shadows, terrain, and mixed vegetation, making traditional pixel-based or threshold-based methods inaccurate. Therefore, we adopted random forest and U-Net models, which can integrate spectral and spatial features (including vegetation index and texture indicators), enabling more robust and general classification in two periods (2012 and 2021). The spectral and spatial integration approach has higher sensitivity to spatial structure and reduces classification errors in complex areas such as transition zones or shadow areas.

It appears to me that very typical patterns of “shade” and “forest” are produced, that should allow a ML model to recognize forest at a somewhat larger scale than at the pixel level. Of course, if you train your model with a pixel-level classification of spectral signatures, the ML model is going to reproduce this, but if you would use real ground-truth data or manually labelled forest areas to train the model, it may be able to really recognize forest and to make use of the shadow rather than to have it only as a nuisance (it will still be a nuisance where whole hillslopes are in shadow, but within the forest, it could become part of the signal, and should be capturable in the texture variables, perhaps if you use a different scale (number of neighbours) to calculate the texture.

Response: Thank you for the comment. We noted that shadows are often associated with canopy gaps in continuous forest areas. In our study area, shadows also occur near adjacent scree slopes, representing terrain characteristics.

Based on observations of orthorectified aerial imagery and field surveys, we would like to classify shadows as a separate class.

Another important piece of information that you need to elaborate upon is how you defined the treeline/ treeline ecotone, in the field and on the classified images, what the survey data consist of, and how you compared the survey data to the treeline location suggested by the classification.

Response: Thank you for the comment. We define the treeline ecotone not as a fixed linear boundary but as a transitional zone where krummholz, such as Yushan Juniper (*Juniperus morrisonicola*) and Yushan rhododendron (*Rhododendron pseudochrystam*), begin to appear within the alpine talus slope (Liao. 2016; Liao et al., 2023). (L89-90).

Reference:

Liao, M. C. Vegetation Structure of Subalpine Ecosystem in Taiwan: A Case Study of Xue Mountain. Unpublished doctoral dissertation. National Chung Hsing University. 2016

Liao, M. C., Wang, W., and Tzeng H. Y.: Study of the Structure and Competitive Coexistence of Subalpine Krummholz Species in Taiwan. Taiwan J. For. Sci., 38(3), 203-220, https://doi.org/10.7075/TJFS.202309_38(3).0002, 2023.

Usually, the accuracy of ML models is much higher within the image it was trained on that on new images, since these may differ in e.g. lighting, season, angle, etc. Figure 7 a and c show that indeed the lighting seems to have been quite different in the two images. Therefore, it is not unlikely, that the classification accuracy was a lot lower for the year not used for model training, so that part of the differences through time could be due to this different accuracy. Did you validate the accuracy for the other year in any way? It would be important to do and show a detailed manual (not based on another automated classification method) validation of the results, seeing whether and where the forest, in particular, is classified correctly, especially at the boundary between forest and non-forest, i.e. in the treeline ecotone. You could e.g. use imagery from Google Maps or Bing, which have a higher spatial resolution for many parts of the world than Worldview images and are often available for past dates, or perhaps there are aerial images available from more local sources. These would allow you to check whether the areas where you detected change really appear to have changed in reality. I guess your field survey data might also show this, but it is currently unclear from the manuscript how these were used. Related to this, especially if you define forest elevational shift by the single highest point, there is a reasonable chance that a change in that point does not represent a real shift. Perhaps you can think of a different, more robust, measure of treeline-ecotone elevation?

Response: Thank you for the comments. We examined images from Google Earth Pro; however, we found that the inter-annual differences in our study area were often subtle and unclear. Therefore, we relied on high-resolution satellite imagery (WorldView) to investigate ATE change in detail.

Our ground truth data were independently generated through manual digitization and expert visual interpretation of the satellite imagery, further cross-validated through proximity to field survey data. These ground truth data were used for training and validating the classification results, ensuring independence from the model and robustness in assessing accuracy.

As mentioned, we attempted to utilize Google Earth Pro and other high-resolution base layers for visual validation. However, due to limitations in temporal availability and clarity for our study area, the above-mentioned sources were not reliable for identifying changes. We have clarified this limitation in the revised manuscript and emphasize our reliance on consistent and high-resolution satellite data.

We appreciate your suggestion to adopt a more robust measure for treeline elevation rather than relying solely on the single highest forest pixel. In response, we applied a bootstrap resampling method (5,000 iterations) to estimate the distribution of treeline elevations. This approach allowed us to assess the statistically robust mean elevation change. Our analysis shows that the treeline has shifted upward by 32.00 meters from 2012 to 2021, with a 95% confidence interval of ±4.00 meters. We revised the manuscript by adding one figure (Figure 9). (L291-293, 309-312).

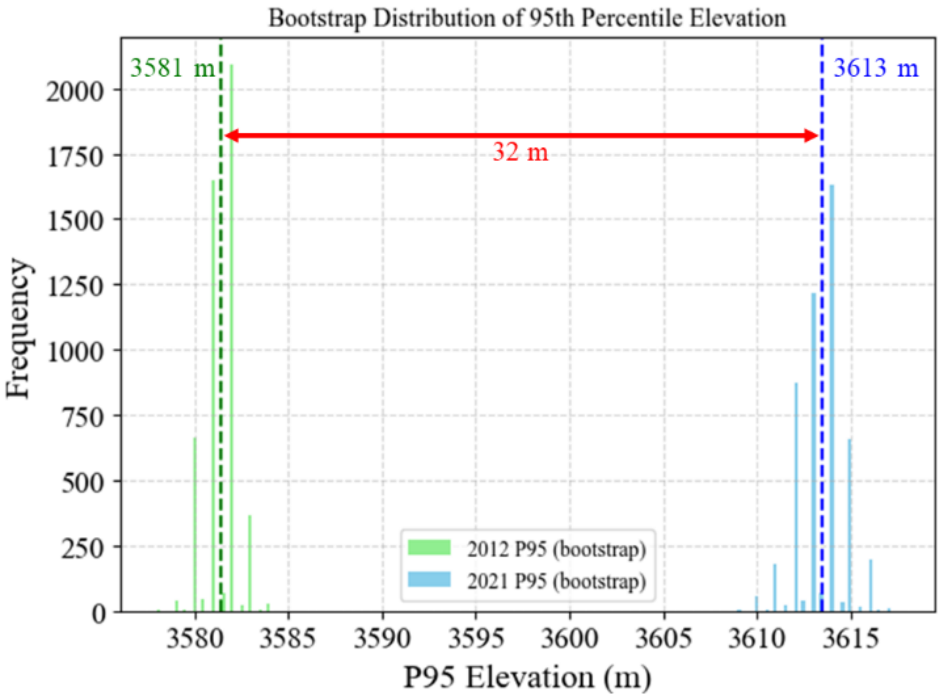

Figure 9. Bootstrap distribution of the 95th percentile elevation of forest cover for 2012 and 2021. The histogram shows the frequency of estimated 95th percentile elevations (P95) based on resampling. Green bars represent 2012 estimates, while blue bars represent 2021. The dashed vertical lines indicate the mean P95 value for each year.

About the presentation: the figures are well-prepared. I like figure 2, in particular, as it nicely explains the different steps taken in the research and the connection between them. However, the captions for all of the figures and tables are much too short. They do not explain what is in the respective figure or table. Please expand and try to make the figures and tables self-explanatory, i.e. if a reader goes to look at them without having read the main text, they can sort of understand what they are about. See my example below for Fig 1. I also would advise not to use title font (capitalized words) in the figure captions or section titles.

Response: Thank you for your comments. We revised all captions for all figures and tables in the revised manuscript.

Some more detailed comments:

You sometimes report more decimal numbers than is reasonable or useful (e.g. L 21, 24, 48, 49) Please check for this and reduce the unnecessary precision

Response: Thank you for the comment. We have revised the numerical values to less than three decimal places in the manuscript.

Avoid sentences like "The F1-score results are shown in Fig. 4." – instead, give the results and then just cite the figure: The result was X (Fig, 4).

Response: Thank you for the comment. We have modified the sentences accordingly in the revised manuscript. (L225).

Title: Change Alpine to alpine (the capital letter would suggest that you worked in the Alps, whereas the small a is used for alpine as a life zone in general)

Title: on Xue Mountain in Taiwan / in the Xue Mountains of Taiwan

Response: Thank you for the comment. The title was modified as follows:

> "Machine learning-based alpine treeline ecotone detection on Xue Mountain in Taiwan"

Abstract:

L14 do not use capitals for alpine treeline ecotone

L18 remove "the" before Random Forest (and not sure that random forest needs to be capitalized)

L19 & 26-27 either use the introduced abbreviation (ATE) or the full term (alpine treeline ecotone), but not both every time

Response (L14, L18, L19 & 26-27): Thank you for the comment. We revised related contents.

Introduction

L33-34 "Alpine zone ecosystems are susceptible to environmental changes compared to other regions": are they really? Maybe they are not, because of the high environmental heterogeneity and small migration distances, compared to latitudinal climate gradients…

Response: Thank you for the comment. We revised the manuscript as follows (L33-35):

> “Alpine zone ecosystems are particularly vulnerable to environmental change due to their high environmental heterogeneity and limited species migration distances, especially when compared to broader latitudinal climate gradients and more resilient lowland regions.”

L35-36 I would suggest referring to this transition as the alpine treeline ecotone (not alpine treeline and not capitalized). This is partly a matter of habit and taste, but it was recently suggested to reserve “treeline” for the climatic potential, and treeline ecotone to the actual observed transition from forest to alpine (or upper forest limit if it is unclear whether the transition is even related to climatic limitations) See e.g. Körner & Hoch 2023 and Malanson 2024. If you decide to follow this terminology, check its use throughout the manuscript.

Response: Thank you for the comment. We want to follow the terminology with the alpine treeline ecotone (ATE) and revise the manuscript accordingly.

L38: Bader et al., 2020 should be 2021

Response: Thank you for the comment. We have corrected it. (L40).

L38-39 “Furthermore, ATE changes illustrate the impact of climate change…” Why “furtherore”? And how do you know that climate change is the driver of change? Are there no other potential drivers?

Response: Thank you for the comment. We totally agree with you that climate change is a major driving force behind changes in alpine treeline ecotone (ATE), and many researches show that there are other potential factors that also influence changes in treeline location and structure. We rephrased this sentence as follows (L40-44):

> “Based on many studies, changes in the alpine treeline ecotone (ATE) illustrate the impacts of climate change on mountain ecosystems, such as the upward migration of tree species and increased tree density. However, these shifts are also influenced by other drivers, including land-use history, altered disturbance regimes (e.g., fire disturbance), herbivory pressure, and species-specific physiological traits (Wang et al., 2016; Johnson et al., 2017; Du et al., 2018; Mohapatra et al., 2019).”

L42 “to study alpine treelines.” You could cite Garbarino et al 2023 here

L41-54 This paragrpahs gives some examples, but the seem a bit of a random pick. Can you highlight how they are somehow connected (e.g. three examples of studies at different spatial resolutions (please provide the sensor resolution for each data source used), three examples of change detection, or something else.

L52-53 These percentage have no meaning, since we do not know what the reverence area was.

L54 Careful, these studies do not tell us anything about the reliability of the methods applied. Maybe the confirm the usefulness or the great potential or something like that, but not the reliability

Response (L42, L41-54, L52-53, L54): Thank you for the comment. We reorganized the content by grouping the cited studies according to spatial resolution and sensor type, clarified the reference areas for percentage changes where applicable, and revised the concluding sentence to avoid overstating the reliability of remote sensing methods. These changes aimed to improve both scientific accuracy and narrative coherence. (L45-60).

L56 “favorable results” of what? Classification accuracy?

Response: Thank you for the comment. We revised “vaforable results” to “promising classification results”. (L62).

L58 municipality

Response: We corrected the spelling, thank you.

L62 the tree Cecropia hololeuca, which has a optically striking shape and colour (I think, check the original paper to confirm)

Response: The tree species *Cecropia hololeuca* exhibits distinctive spectral signatures associated with its large leaves and crown structure. Furthermore, due to the abundance of *Cecropia hololeuca*, their visually prominent shape and bright gray coloration make them readily identifiable to the human eye in RGB imagery.

Reference:

Wagner, F. H., Sanchez, A., Tarabalka, Y., Lotte, R. G., Ferreira, M. P., Aidar, M. P. M., Gloor, E., Phillips, O. L., and Aragão, L. E. O. C.: Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images. Remote Sens. Ecol. Conserv., 5(4), 360-375, https://doi.org/10.1002/rse2.111, 2019.

L63 “The classification accuracy for Cecropia hololeuca species reached 97%, with an IoU of 0.86.” This is a bit too much technical detail at this point.

L68-69 “Based on these studies, applying…” à Based on these studies, we conclude that applying…

Response (L63, L68-69): Thank you for the comment. We revised these sentences. (L67, L71-72).

L80 with “studies on the volume estimation”, do you mean “studies estimating wood volumes”?

L81 It is not clear here whether these studies were done in the alpine treeline ecotone, or in the forest below.

Response (L80, L81): Thank you for the comment. We revised these sentences as follows (L84-87):

> “Most ecological studies conducted in this research area have focused on Taiwan fir forests, with several researchers estimating wood volumes, competitive pressure, forest structure, and spatial distribution of the species primarily through field surveys conducted below the alpine treeline ecotone.”

L82 remove the sentence referring to the figure (just add (Fig 1). Instead please explain here what “krummholz” means to you. How do you define it? And the same for “forest” This is quite relevant in a treeline context.

Response: Thank you for the comment. We defined krummholz and revised the manuscript as follows (L193-197):

> “Ground truth data for the study area were manually labeled using a pixel-based approach and categorized into four classes: (1) Bare land, referring to areas of exposed soil, rock surfaces, or sparsely vegetated ground; (2) Forest, defined as regions with dense, continuous tree canopy cover; (3)

Krummholz, representing stunted, shrub-like trees typically found at high elevations near the treeline and shaped by wind or snow pressure (Liao et al., 2023); and (4) Shadow, representing regions with low reflectance caused by topographic shading or solar angle effects."

Reference:

Liao, M. C. Wang, W., and Tzeng H. Y.: Study of the Structure and Competitive Coexistence of Subalpine Krummholz Species in Taiwan. Taiwan J. For. Sci., 38(3), 203-220, https://doi.org/10.7075/TJFS.202309_38(3).0002, 2023.

Figure 1 Please provide a self-explanatory caption. E.g. "Location of the treeline ecotone study area in the Xue Mountain glacial cirques in Shei-Pa National Park (top-right map) in north-central Taiwan (top-left map). The red marker in the aerial image (bottom-left map) indicates….. The digital elevation model shown in the bottom-right image shows the same area as the aerial image and covers the entire study area" Instead of having to refer to the "top-left map" you could also add letters to each panel.

Thank you for the comment. We modified Figure 1. (L91-95).

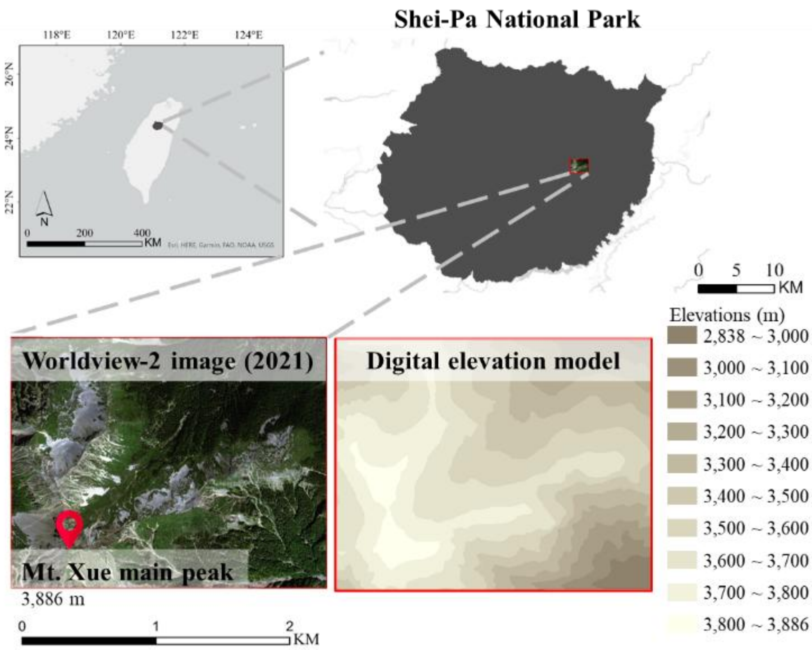

Figure 1. Geographic location of the treeline ecotone study area in the Xue Mountain glacial cirques in Shei-Pa National Park (top-right map) in north-central Taiwan (top-left map). The red marker in the Worldview-2 image (bottom-left map) indicates the research area. The digital elevation model shown in the bottom-right image shows the same area as the Worldview-2 image and covers the entire study area"

Methods & **results**: please write the methods and results sections in the past tense

Response: Thank you for the comment. We revised the methods and results sections in the past tense.

L89 How was the optimal model selected, did you validate the classification at all? How did you obtain the "ground truth"- from the flow chart fig 2, it looks like you obtained it from the image itself, rather than from survey data? Survey data are not in the flow chart at all.

Response: Thank you for the comment. We have modified Figure 2 to illustrate the ground truth. The ground truth data were obtained from GPS survey points and then manually labelled with expert interpretations.

L93 I suggest writing "is reported to be within 3 meters"

Response(L89, L93): Thank you for the comment. We rephrased the sentence. (L111).

Figure 2: This is a nice figure (but the caption needs to explain what this is a research flow for (e.g. "Research flow for classifying Worldview images of a treeline ecotone on Mt Xue in Taiwan for detecting changes in forest cover."). A detailed question: in section 7, the two images that are subtracted indeed look a bit different, but in Figure 6, they look much more similar. What explains this discrepancy? Maybe in 6, b and c are accidentally the same image..?

Response: Thank you for the helpful comment. We modified the caption of Figure 2 as follows (L102-106):

> "Figure 2. Research flow for classifying WorldView-2 images of a treeline ecotone on Mt. Xue in Taiwan to detect treeline changes. The process begins with WorldView-2 satellite image acquisition, followed by feature extraction (spectral bands, vegetation indices, and texture features), model training using Random Forest (RF) and U-Net, accuracy evaluation, feature selection, and temporal analysis of alpine treeline changes between 2012 and 2021."

Regarding Figure 6, we appreciate your comments. After careful re-examination, we confirm that Figure 6(b) and Figure 6(c) are not the same image. The differences are subtle but can be observed, for example, changes in the forest area in the northwest and the shadowed region in the northeast. Specifically, Figure 6(b) shows the classification results using all 77 features, while Figure 6(c) presents the results after applying feature selection, which reduces the number of features to 61. The overall appearance is similar, but minor differences remain in certain regions.

L104 "and GPS was used to record survey points" This addition seems out of place…

Response: Thank you for the comment. We revised this sentence as follows (L123-124):

> "GPS devices were used to record field survey points, which were subsequently used to verify ATE positions and assist in manual ground truth labeling"

L107 2.4 Vegetation indices

Response: Thank you for the comment. We modified this sentence. (L127).

L115-116 This reads like it came out of the research proposal. Here, please use the past tense.

Response: Thank you for the comment. We revised these sentences. (L134-135).

L149 "calculated using the following formula:" and then no formula follows…

Response: Thank you for the comment. We revised this sentence. (L168-174).

L173, Fig 3 With “ground truth data” you mean manually classified images for training and validation? That is not really ground truth, is it? Maybe call this “labelled data” instead?

Response: Thank you for the comment. In this study, based on the GPS-based field survey points and expert interpretation, we manually labeled images as "ground truth." The labeled ground truth data were used to serve as the reference for both training and validation. In remote sensing and image classification literature, ground truth can refer to the manually labeled data used for validation and accuracy assessment.

L175 Can you define these classes? For example what is “bare land”? Rocks? Soil? Does it include alpine vegetation other than krummholz?

Response: Thank you for the comment. We defined and revised the manuscript as follows (L193-197):

> “Ground truth data for the study area were manually labeled using a pixel-based approach and categorized into four classes: (1) Bare land, referring to areas of exposed soil, rock surfaces, or sparsely vegetated ground; (2) Forest, defined as regions with dense, continuous tree canopy cover; (3) Krummholz, representing stunted, shrub-like trees typically found at high elevations near the treeline and shaped by wind or snow pressure (Liao et al., 2023); and (4) Shadow, representing regions with low reflectance caused by topographic shading or solar angle effects.”

L176 “The dataset was randomly split, with 80% used for training and validation and 75% and 25% allocated for training and validation,”… something appears to be wrong here.

Response: Thank you for the comment. The dataset was randomly split. We clarified it in the revision as follows (L202-204):

> “Specifically, 80% of the patches were randomly selected for training and validation (with a 75/25 split), and the remaining 20% were used as an independent test set. In total, 66 patches were used for training, 22 for validation, and 22 for testing.”

L194 Explain what the F1 score is, E.g. the accuracy of the models can be depicted by the F1 score, which exceeded… etc.

Response: Thank you for the comment. We revised this sentence. (L223-225).

> “The F1-scores, representing the harmonic mean of precision and recall, provided a balanced assessment of classification performance. All classes achieved F1-scores above 0.6 (Fig. 4). Forest and krummholz were more frequently misclassified with one another due to their similar vegetation structures, while bare land and shadow were more easily distinguished, achieving F1-scores above 0.8.”

L195 what do you mean with “they tend to influence each other more”? That they are confused more often?

Response: Thank you for the comment. We revised this sentence as follows (225-227):

> "Forest and krummholz were more frequently misclassified with one another due to their similar vegetation structures, while bare land and shadow were more easily distinguished, achieving F1-scores above 0.8."

L199-203 It seems that you may be over-interpreting the differences in F1 scores between the combinations. Most differences seem quite minimal. I would suggest writing that the combinations performed similarly well, pointing only at the larger differences that may actually mean something in terms of model performance (for the U-net model). It would also help to remind the reader what features are included in what combination, as this teaches us something about the importance of e.g. texture for recognizing different cover classes.

Response: Thank you for the comment. We revised this paragraph as follows (L228-235):

> "Overall, the different feature combinations produced similar classification performance, with only minor differences observed across classes and models. In the RF model, bare land and shadow achieved the highest F1-scores (0.905 and 0.866, respectively) when using Combination 1 (spectral bands only). Forest and krummholz performed slightly better with Combination 4 (spectral bands, vegetation indices, and texture features), achieving F1-scores of 0.827 and 0.776, respectively. In the U-Net model, Combination 1 yielded the best result for bare land (F1 = 0.889), while Combination 4 slightly improved the classification of forest (0.828), krummholz (0.886), and shadow (0.869). These findings suggested that incorporating vegetation indices and texture features improved model performance for specific vegetation classes, particularly in the U-Net model, although overall improvements remained relatively modest."

L204 You just presented some accuracy metrics, and here you suddenly present other accuracy metrics; that is a bit confusing. Maybe ad "overall accuracy" (as opposed to the class-wise F1 scores), but I recommend just presenting the results here ("Similar to the accuracy patterns for the individual classes, in the U-Net models, the overall classification accuracy improved as the number of features increased, whereas for the RF model this was not the case (Table 5)"), not the table as such.

Response: Thank you for the comment. We revised these sentences as follows (L236-238):

> "The overall accuracy (OA) and Kappa coefficient for each feature combination were summarized in Table 5. Similar to the accuracy patterns for the individual classes, in the U-Net models, the OA improved as the number of features increased, whereas for the RF models this was not the case (Table 5)."

Figure 4 Pleas repeat here what the different feature combinations were

Response: Thank you for the comment. We redrew Figure 4. (P12).

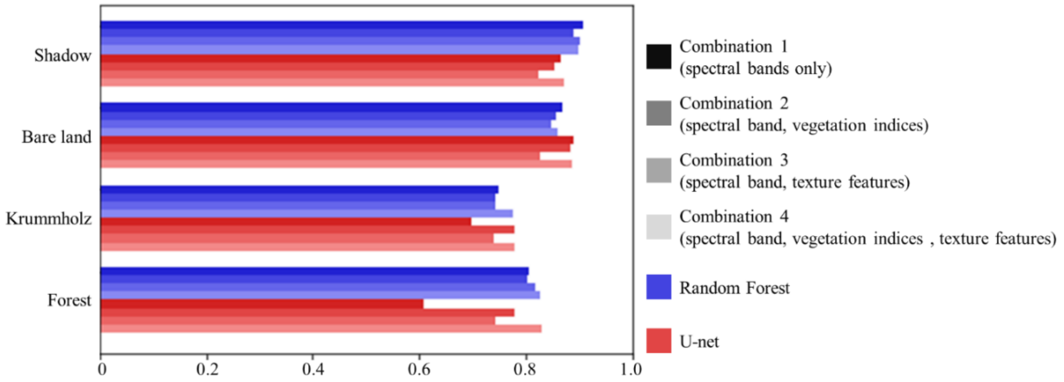

Figure 4. F1-scores for four land cover classes (forest, krummholz, bare land, shadow) using RF and U-Net models with different feature combinations."

Table 5 Explain abbreviations and combinations and number in brackets and bold font in the caption

Response: Thank you for the comment. We revised the Table 5 as follows (L246-249):

> "Table 5. Evaluation of classification accuracy using different feature combinations and models. Overall accuracy (OA) and Kappa coefficient are shown for Random Forest (RF) and U-Net models. Numbers in parentheses indicate the number of input features. Bold values indicate the best results for each metric."

L214 "The treeline is determined based on the boundary between bare land, forests, and krummholz." – I suggest using "The upper limit of the treeline ecotone", and explaining better how you define this line, since now there are three classes, and those will all three be heterogeneously spread across the landscape….Since treeline is defined by trees and the ecotone is generally also described by the patterns of tree cover, the accuracy of forest detection seems to be the most important, so in line L221 you could add behind "texture features were relatively less important" ", as also suggested by the low F1 scores for combination 3 (spectral + texture; Fig 4), although for forest, in particular, texture strongly increased the classification accuracy relative to providing just spectral information, but vegetation indices increased it more".

Response: Thank you for the comment. We revised this paragraph as follows (L251-252, 261-264):

> "The upper limit of the ATE was determined based on the spatial distribution boundary where patches of forest transitioned into krummholz and bare land."

> "In contrast, texture features were relatively less important, as also suggested by the low F1 scores for combination 3 (spectral, texture; Fig. 4). However, for the forest class in particular, texture features significantly improved classification accuracy compared to using only spectral bands, and the inclusion of vegetation indices contributed even more to the performance."

L218 remove the sentence that introduces Fig 5, just add (Fig 5)

Response: Thank you for the comment. We revised this sentence. (L258-259).

L221 remove the line break, it does not look like a new paragraphs should start here. If anything, start it with Using these 61…

Response: Thank you for the comment. We removed the line break.

L222 Why is it important to reduce training time? Is it more important to reduce training time than to get a better model fit? I guess this could become important if one would want to apply the model more broadly, but for your particular application it does not appear to matter. On the other hand, there is the concept of parsimony in model selection, i.e. select the model with the best fit, but penalizing for model complexity, i.e. make the model as complex as necessary, but no more. Is this concept related, does it apply to machine-learning models? If the last features add 5% of accuracy, that is the same order of magnitude as you effect size (the change in forest cover), so those 5% may be relevant…? You see, I am a bit confused. Perhaps it needs a bit more explanation why you decided to use feature selection.

Response: Thank you for the comment. Reducing the number of features not only shortened the computation time but also improved performance. After feature selection, the number of features was reduced from 77 to 61. The classification accuracy on forest classification was stable, but the classification accuracy for krummholz improved by 2%, which is essential for ATE interpretation. This follows the principle of parsimony, keeping the model simple while maintaining prediction capability.

L223 Remove the "additionally. That difference is not a difference…

Response: Thank you for the comment. We have revised this sentence. (L268-269).

Table 6 Could you provide the training time in hours, so the it is easier to understand the order of magnitude?

Response: Thank you for the comment. We have revised the table 6. (P13).

Figure 6 These maps are not so informative. Maybe one bigger map with the classes and including missclassifications would show better how good the classification is. The "Ground truth" looks like another automatic classification, please explain well in the methods how you got to this map, and maybe call it something other than "ground truth"…

Response: Thank you for the comments. The "ground truth" labels in our study were not derived from automated classification but from manual image digitization confirmed by expert field surveys. Additionally, we redrew Figure 6 to better highlight the classification results and their differences, making the comparison more informative.

Figure 7 Please print as big as possible, it is very hard to see anything on these images. Could you draw the boxes shown in b and d in the images in a and c, and remove the big white boxes with 1 and 2 (they block the view)? Also, the symbols for the field investigations are hard to see. It would also be more informative if not only the location, but also the vegetation types of the field survey points would be shown. As it is , it is unclear what the field survey data are. Please explain these data in the methods section and again briefly in the figure caption and in line 234-235. In any case, to be able to evaluate the fit, the images would need to be much bigger. You could also plot the fits of the field (real ground truth!) and the image classification in a separate graph.

Response: Thank you for the comment. We revised the Figure 7 and paragraph as follows (P15, L283-286):

> "A U-Net model was trained using 61 selected features derived based on feature importance. The trained model was applied to classify satellite images from 2012 and 2021. The classification results were validated against field survey data collected in 2021, which recorded vegetation types and the position of the tree line along an elevational gradient. As shown in Fig. 7, the tree line derived from the classification closely aligns well with the tree line identified through GPS-based field survey points."

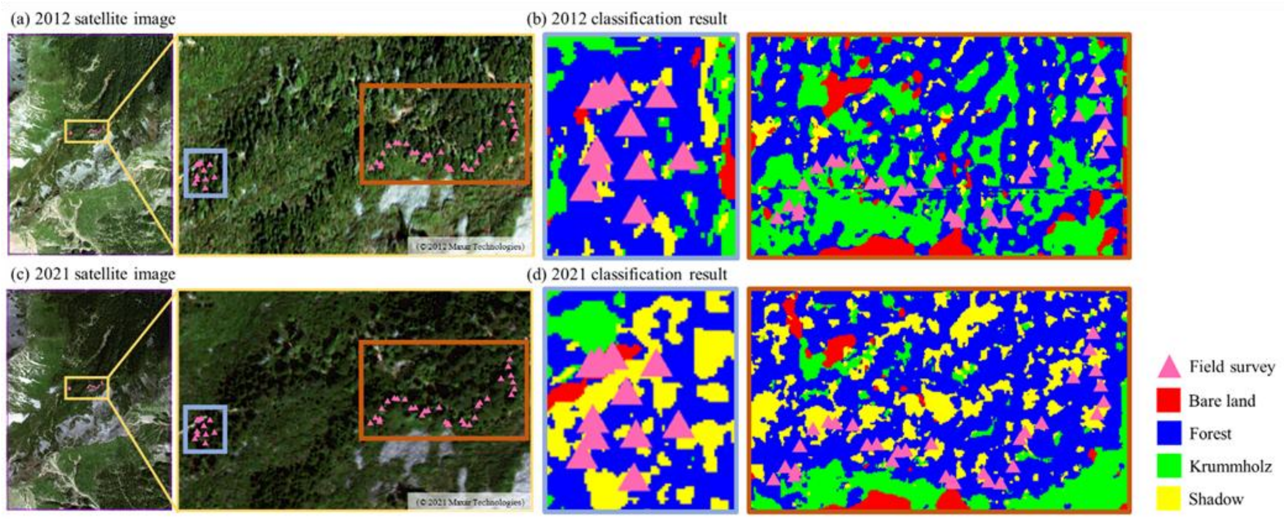

Figure 7 Comparison of satellite imagery and classification results from 2012 and 2021. Panels (a) and (c) show high-resolution satellite images for 2012 and 2021, respectively. Colored boxes in these images indicate the enlarged areas shown in (b) and (d). Panels (b) and (d) present the classification results of the corresponding enlarged regions using a U-Net model trained with 61 selected features. Triangles mark GPS-based field survey points.

L232 Decadal changes in the treeline ecotone

Response: Thank you for the comment. We revised this subtitle. (L282).

L235 avoid using tree line, stick to treeline or, even better treeline ecotone

L235 Since you have not explained how you define treeline, this statement is impossible to follow.

Response(L235): Thank you for the comment. We revised the manuscript, and we defined the treeline ecotone as follows (L89-90):

> "We define the treeline ecotone not as a fixed linear boundary but as a transitional zone where krummholz, such as Yushan Juniper and Yushan rhododendron, begin to appear within the alpine talus slope."

L236 "Over a decade, the proportion of forest and shadow areas increased by 3.4% and 8.5%,…." Really? The proportion of shadow area increased in the last decade?? Obviously this is just a matter of lighting when the image was taken, so I recommend rephrasing this result.

Response: Thank you for your comment. Our classification results show that forest and shadow areas increased by 3.4% and 8.5%, respectively. Considering that shadow areas can be affected by lighting conditions at the time of image acquisition, the observed increase may not fully reflect actual land cover changes. In future work, we will further investigate the impact of lighting on classification results and refine our processing methods accordingly. We revised the manuscript as follows (L287-289):

> "Over the decade, the proportion of forest area increased by 3.4%, indicating a trend of forest expansion. Meanwhile, the proportion of shadow area also increased by 8.5%; however, this is likely due to differences in lighting conditions and satellite viewing angles between the 2012 and 2021 image acquisitions rather than an actual ecological change."

L239 how did you define the elevation distribution of the forest? The uppermost forest pixels?

Response: Thank you for this thoughtful question. To determine the elevation distribution of forests, we extracted the elevation values of all pixels classified as forests from the DEM. Instead of simply using the uppermost forest pixels, we defined the upper elevation limit as the 95th percentile of forest pixel elevations. To robustly estimate treeline elevation change and its uncertainty, we also used a bootstrapped resampling approach to calculate 95% confidence intervals.

Figure 8 caption: between 2012 and 2021. It would also be helpful to see the persistent forest cover here.

Response: Thank you for the comment. We modified Figure 8 as follows (P16):

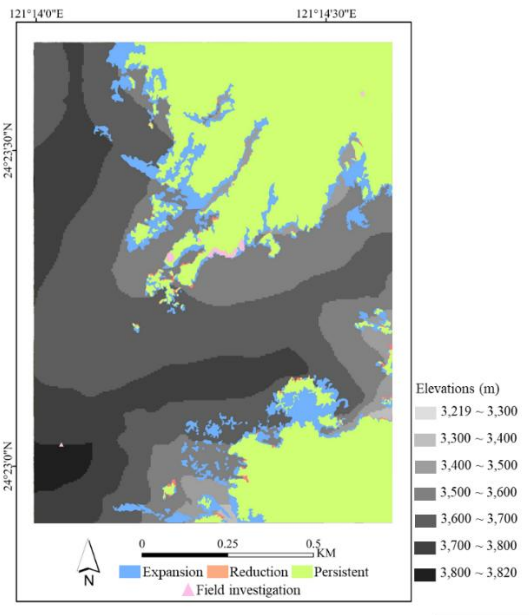

Figure 8. The spatial distribution of forest area changes from 2012 to 2021. Forest expansion is marked in blue, reduction is marked in orange, and persistent is marked in green.

L238, Table 9: it may be better to express the area changes in e.g. ha, instead of km2, to get nicer numbers.

Response: Thank you for the comment. Table 9 has been updated to express area changes in hectares (ha) instead of square kilometers (km²), resulting in more intuitive values for the reader. (P17).

L240-241 “with the most significant changes occurring in the 3,500 to 3,600 m range.” If this is where most of the treeline ecotone was, it would be worth mentioning this here

Response: Thank you for the comment. We revised the sentence as follows (L294-295):

> “With the most significant changes occurring in the 3,500 to 3,600 m range, which corresponds to the primary treeline ecotone change zone in the Xue Mountain region.”

L241 and/or L308 “In comparison, the most stable area was observed in the 3,700 to 3,800 m range.” – explain here that hardly any forest was found here at any time.

Response: Thank you for the comment. We revised the sentence as follows (L296-297):

> “In comparison, the most stable area was observed in the 3,700 to 3,800 m range, where minimal forest presence was detected in both 2012 and 2021, reflecting physiological limits of trees.”

Table 7: please also provide the forest area in each belt in 2012 and the % change

Response: Thank you for your comment. We have revised Table 9 as suggested. However, since we have deleted the original Table 8, the current Table 9 is now renumbered as Table 8. Table 8 as follows:

Table 8. Forest area, expansion, and reduction across different elevation from 2012 to 2021. The table includes forest area in 2012, net changes in area, and corresponding percentage changes.”

Elevations (m)	Forest Area in 2012 (ha)	Expansion area (ha)	Reduction area (ha)	Net Change (ha)	Change (%)
3300~3400	6.99	0.28	0.03	0.25	3.6
3400~3500	12.43	2.21	0.08	2.13	17.1
3500~3600	8.40	5.10	0.23	4.87	58.0
3600~3700	3.26	2.88	0.06	2.82	86.4
3700~3800	0.78	0.02	0.00	0.02	2.5

Discussion

L256-267 & 275-288 As a general advice, better start with you results and then contrast or align it with other studies, rather than the other way around. Now these paragraphs read a bit like a second introduction, again listing studies without any obvious logical connection between them. If you start with your results, you can use connections like “In contrast…”, or “Likewise” to make a clearer connection between other people’s findings and your results.

Response: Thank you for the comment. We reorganized the discussion sections as follows (L319-329, 342-346):

> “Our findings reveal that, from 2012 to 2021, the alpine treeline ecotone (ATE) in the Xue Mountain glacial cirque experienced an upward shift of 32.00 ± 4.00 meters, along with a pronounced densification

> of forest cover. This finding aligns with patterns observed in other mountainous regions worldwide. For example, in Taiwan's Hehuan Mountain and Yushan, similar upward shifts in treeline position and increases in forest density have been reported (Greenwood et al., 2014; Chung et al., 2021). Likewise, Davis et al. (2020) observed an upslope advance of 0.83 ± 0.67 m/year for several tree species in the Rocky Mountains of Canada. In contrast, studies in the European Alps have noted significant reductions in snow cover and increased alpine vegetation productivity, potentially enhancing local carbon sequestration, although with a limited global impact (Rumpf et al., 2022). Additionally, in the eastern Himalayas, over 80% of trees have already reached the thermal treeline, with projected upslope migration of 140 meters by the end of the 21st century due to warming (Wang et al., 2022). These comparisons support the robustness of our observed treeline dynamics and highlight both global consistency and regional variation in alpine ecosystems response to climate change."

> "While OA was used as the primary selection criterion, we also confirmed that these top-ranked features maintained or improved F1-scores for the forest class, which is the primary concern in detecting treeline changes. We recognize that the process of optimizing OA values may sometimes overlook minority or ecologically important classes. Therefore, we specifically examined the F1-score for the forest class—our primary concern for treeline detection—and verified that its classification performance was not compromised."

L271 Was there any relationship between the treeline pattern and the local change? E.g. did abrupt treeline stay more stable than ones with tree islands? A bit more ecological interpretation of the patterns found would be interesting (e.g. relationship with topography).

Response: Thank you for the comment. Yes, based on our field surveys, there are topographic correlations between treeline patterns and local changes, which we are currently working on in another manuscript.

L275-288 Can you discuss here what the difference is between feature selectin based on the overall accuracy and feature selection based on the accuracy of your target land-cover class (which was obviously not shadow, but forest?

Response: Thank you for the comment. We revised the paragraph as follows (L342-346):

> "While OA was used as the primary selection criterion, we also confirmed that these top-ranked features maintained or improved F1-scores for the forest class, which is the primary concern in detecting treeline changes. We recognize that the process of optimizing OA values may sometimes overlook minority or ecologically important classes. Therefore, we specifically examined the F1-score for the forest class—our primary concern for treeline detection—and verified that its classification performance was not compromised."

L292 Maybe do not mention shadow here, since that is more a no-data area than a land-cover class…

L294-295 These numbers are very technical for a conclusion section…

L296 "SEVI, Y, B, G, and NDVI2." Here, since readers may read the conclusions without having read the whole paper, you might want to explain the abbreviations.

L297 Again, this cannot be understood without an explanation about what the survey data are and how treeline was defined.

L299 denser or expanded?

L299 at higher elevations

Response (L292, L294-295, L296, L297, L299): Thank you for the comment. We revised the paragraph as follows (L364-375):

> "This study investigates changes in the ATE of the Xue Mountain glacial cirques in Taiwan from 2012 to 2021, utilizing WorldView-2 imagery in conjunction with Random Forest and U-Net models. By incorporating spectral bands, vegetation indices, and texture features, we achieved improved classification accuracy and computational efficiency. Feature selection identified the most important variables as the Shadow-Eliminated Vegetation Index (SEVI), Yellow (Y), Blue (B), Green (G) bands, and Normalized Difference Vegetation Index (NDVI2). The treeline was defined not as a fixed linear boundary but as a transitional ecotone where krummholz species—such as Yushan juniper (*Juniperus morrisonicola*) and Yushan rhododendron (*Rhododendron pseudochrysanthum*)—begin to appear within the alpine talus slope. This delineation was based on both satellite classification results and GPS-referenced field survey data. Over the past decade, forest cover in the study area expanded by approximately 0.101 km^2, indicating both denser canopy growth and outward expansion. In addition, the upper limit of forest distribution rose by 32.00 ± 4.00 meters, indicating an upslope shift of the treeline at higher elevations. These findings provide new insights into treeline dynamics in Taiwan's alpine environment and demonstrate the potential of high-resolution satellite imagery for long-term ecological monitoring."

References: please ident them (the hanging parts of each reference, rather than the main author name) to make them easier to navigate

Response: Thank you for the comment. We revised the manuscript.

References cited

Garbarino, M., D. Morresi, N. Anselmetto, and P. J. Weisberg. 2023. Treeline remote sensing: from tracking treeline shifts to multi-dimensional monitoring of ecotonal change. Remote Sensing in Ecology and Conservation **9**:729-742. https://doi.org/10.1002/rse2.351

Körner, C., and G. Hoch. 2023. Not every high-latitude or high-elevation forest edge is a treeline. Journal of Biogeography **50**:838-845. https://doi.org/10.1111/jbi.14593

Malanson, G. P. 2024. Inclusions and exclusions in treeline definitions. Journal of Biogeography **51**:54-56. https://doi.org/10.1111/jbi.14729