Reviewer 1 Comments

General comments

The manuscript is well written, structured, and presents its findings in a logical and accessible manner. It addresses an important and timely topic within the field. The authors employ an innovative deep learning approach to identify treeline tree species based on high-resolution satellite images. The methodology is comprehensible and appropriate. The results are well interpreted, while giving them a stronger ecological context and linking it stronger to the scientific literature in the discussion would be beneficial. Overall, the study makes a meaningful contribution when some major points are addressed.

We thank the reviewer for these favorable comments! We are glad they think the manuscript will be a useful contribution to the field.

While I greatly appreciate the methodological and technical details, the manuscript lacks reporting results of CNN application in the two study areas. The authors clearly discuss limitations of the CNN application in the discussion section. However, reporting on the segmentation results and limber pine distribution (area cover, patch sizes, elevation) would underline the relevance of the approach, help the readers to think of further application and provide data for further ecological investigation. In L121 the authors state that the current treeline distribution of limber pine is unknown. Addressing this gap with the presented approach, as a proof of concept, would enhance relevance of the manuscript.

We apologize for the confusion and appreciate the opportunity to clarify the manuscript. At this point in time, the proposed segmentation and subsequent object-based classification has not been attempted. We plan to tackle those methods in a subsequent paper building on this work. We outlined the plan in the conclusions to help readers think about applications of the present work. We have rephrased lines 636-639 of the Conclusions to read the following:

"We anticipate that outputs from our CNN models, in the form of pixel-level probability maps for classified species, may be easily processed into segmented images of tree and shrub objects for object-based classification. While beyond the scope of the present work, segmentation and object-based classification will be the necessary next steps to apply these methods at a landscape scale. Once these methods are developed, we may be able to use satellite imagery to map limber pine's distribution at treeline and monitor the distribution of this important species as climate changes."

The manuscript is long and features 12 figures and 3 tables. Please consider the most important display items and move others to the supplementary material.

We have done this but we do consider these figures and tables to be the minimum necessary to communicate the methods and results. We have responded to the reviewer's suggestion below to move the example best hit and worst miss and explained why we believe those two figures are necessary.

The results section has several sections that should be considered for the discussion. The discussion section will benefit from a stronger link to scientific literature.

The discussion section now also cites the literature the reviewer is referring to (documentation of species distributional tendencies as context for the CNN's reliance on elevation as a predictor). We do not wish to move the sections in the results in question because those sections provide timely context for the results synthesis/overview (see below).

With these revisions, the manuscript will make a strong and timely contribution to the field.

We thank the reviewer for their valuable time and feedback.

Specific comments

L33 Please provide a concise definition of the alpine treeline in the context of your manuscript. I furthermore suggest to use elevational instead of altitudinal.

Thank you! We revised the mid-sentence definition to "the elevational limit of tree occurrence in mountain ecosystems". We considered defining the "alpine treeline ecotone" and decided on "treeline" for brevity, given that the primary focus of the paper is methodological.

L85 according to http://dx.doi.org/10.3390/rs12101667 this approach would be classified as instance segmentation?

Good question, and thank you for checking. In this case, Mishra et al. (2018) used spectral difference segmentation—image segmentation vs. instance segmentation in the paper you linked. Please see Section 2.5.1 and Figure 4 of https://www.mdpi.com/2220-9964/7/11/445. Reviewer 2 was also confused by this section and we have revised the section to read, "Mishra et al. (2018) succeeded in achieving 73% overall accuracy in classifying four tree species across a ~ 140 m x 80 m region of the Himalayas using multispectral UAS imagery. This success highlights the potential and effectiveness of high-resolution UAS data for treeline species identification using an object-based classification approach with spectral difference segmentation (a form of semantic segmentation)."

L91 please briefly introduce LIDAR for the less versed readers

We have added a definition of lidar, including the acronym: “Airborne hyperspectral imagery and lidar (or LiDAR—light/laser imaging, detection, and ranging—an active remote sensing platform that uses time lags in reflected laser pulses to measure distances and so map the 3-D surface structure of sensed objects) have also commonly been used in concert for tree species identification…” We have seen the acronym represented as “lidar”, “LIDAR”, and “LiDAR” in the literature.

L124-125 please be more specific on your study areas and add some numbers, e.g. area, elevation range

We have added specifics about the study areas as requested. We summarized the area, elevation range, and the reason for the study area locations (reasonable trail access): “We chose the study areas for their proximity to trail access within the imagery; the Longs Peak study site (0.72 km^2, ranging from 3,250-3,620 m elevation) was a 4-6 km hike from the Longs Peak Trailhead and the Battle Mountain study site (1.37 km^2, ranging from 3,250-3,560 m elevation) was a 6-8 km hike from the Storm Pass Trailhead. Field work was limited to early mornings before afternoon convection developed into storms with lightning hazards, except for the rare clear day. Access was therefore important for obtaining adequate data for training and validation.”

L135 (Fig 1) The figure would profit from some editing and streamlining:

* While the study areas are very mall (they would be of interest to the reader), the general location in N-America takes a lot of space (while adding little to the context).

* A small insert for the general geolocation would be sufficient

* Having RMNP (e.g. with land cover) as a main map would be nice

* Having an insert with the study areas and the WV3 image as insert in detail would help the readers to better understand the local context

* Symbologies are hard to keep apart

* Streamline sources

We have implemented these helpful suggestions, except that the sources were required by Biogeosciences to be put either in the figure or in the caption. This citation is from ESRI; we cannot streamline it. We did move it to the figure itself in small font, and we hope it will be less distracting there. We hope the reviewer finds the revised map and inset maps more helpful. Please see the attached PDF with the revised figure.

L144 please specify how many GCPs where taken. Was there a certain protocol (e.g. with regard to spacing or spatial coverage)? This could be added to the supplements A.

We have added a paragraph to the beginning of Appendix A, as well as a map showing the locations of the GCPs used in orthorectification:

“Collecting ground control points (GCPs) with a multispectral satellite without tasking the satellite (and placing targets) is quite challenging; we selected GCPs opportunistically, aiming to cover as much of the image as possible. Much of the image, especially lower in elevation, was forested, and taller trees obscured ground features. Much of the rest of the image was tundra, with very few targetable features. We collected GCPs at the corners of switchbacks and at trail junctions, always at or above treeline where these features could be clearly seen in the image. We were able to collect 15 GCPs (Figure A1), six of which were in the Battle Mountain study area and four of which were in the Longs Peak study area.”

Please see the attached PDF with the added map.

L150 was there a DEM involved in the process and if yes what were the specs?

We do not have those details. Maxar performed the orthorectification in 2020 and did not provide details on their methodology. We ordered a very small (minimum allowable) order of imagery; we were not able to obtain further information.

L151-L154 please reword and clearly state your choice (nearest neighbor), the dismissed options, and your rationale.

We have now elaborated:

“We requested the nearest neighbor resampling method specifically, to preserve the original radiance values, as opposed to any method that would substitute the original values with a statistical summary or interpolated value. Cubic convolution resampling is commonly used because it results in a smoother image, but it alters the data values, effectively introducing noise into the data. “Pansharpening” is a similarly inappropriate technique for any analysis that relies on data precision; pansharpening artificially improves the spatial resolution of datasets by modelling statistical relationships between coarser-resolution multispectral imagery with finer-resolution panchromatic imagery and interpolating likely values. The result may aid in visual interpretation for humans, but no new real information is being provided to a CNN. Each larger multispectral pixel is a mixture of spectral signatures of objects at the finer resolution; moving to the finer resolution makes assumptions about the relationship between the two datasets that introduce error and distortions into the data (Zhang et al., 2023). The coarser-resolution multispectral data may not have enough information for realistic assumptions to be made, and it is likely this approach would inflate the sample size without adding new information. We chose to use the data in their native resolutions and adjust our CNN architecture to accommodate the several spatial resolutions. This way, we are still allowing the CNN to

learn from the higher-resolution panchromatic data without inflating the sample size of our multispectral data."

L157 please consider rewording the section header

We have attempted to do so for clarity at the expense of some verbosity:
"2.3 Field Collections of Polygons Delineating Species Patches" (vs. the original, "Species Polygon Collections").

L158-160 was there a protocol for the polygon delineation? How was coherence/ separation of patches defined? Please give some basic stats on the patches (e.g. mean, max, min area) and include them in Tab1.

We added this paragraph for clarification: "We delineated species polygons over five field seasons (2019-2023) to obtain the largest sample size possible in the time available. Our haphazard selection (vs. strict randomization) of species patches that met our criteria was also intended to enable us to obtain as many samples as possible. Species patches were delineated if they were 1) greater than approximately 1 m in area and 2) purely a single species with no contamination from co-occurring species. Ideally, patches were selected which were separated from other tree and shrub species by alpine tundra or rock, but this was not always possible. If patches were bordered by other species or if two pure species patches adjoined, we paused collection to walk to the opposing side of the patch before resuming collection, creating a straight line to indicate greater caution during the later labelling process to ensure training pixels were selected which did not contain any of the adjacent species." We also added areal summaries of the region of interest polygons to Table 1.

L164-166 what was the minimum dimensions (height, area ...) to consider a plant to be a tree/shrub in your context?

We added these lines for clarification: "The species selected for classification are abundant at these treeline sites and are representative of dominant tree and shrub species at treeline in RMNP. Two shrub species that are also abundant were excluded because they form smaller patches at treeline with areas below the image resolution (less than 1.2 m): common juniper (*Juniperus communis* L.) and shrubby cinquefoil (*Dasiphora fruticosa* (L.) Rydb.)."

L171-172 have you identified any offset and if yes how much was it and how did you deal with it?

Yes, we have added these lines to clarify this: "A slight offset of approximately 0.5 m was identified, which reduced the number of viable polygons and pixels for inclusion. We used

the shapes of the vegetation patches in the panchromatic imagery and the shapes of the polygons to identify the correct locations of polygons in the imagery and select only pixels that fell within the field-delineated vegetation patches with high confidence. A single operator performed all pixel labelling, and that operator was most often the person in the field delineating patches."

L172-173 had there been other images than from Maxar?

No, we changed this to "the multispectral and panchromatic images" to avoid confusion.

L183 please consider rewording the section header

Please clarify – we do not understand the issue with "Convolutional Neural Network (CNN) Modelling Methods".

L193ff you could put this intro to CNN/overview in an own section as well and introduce it in the paragraph above.

We have done as requested.

L235ff please state clearly which software platform(s) you have used for your CNNs.

We added a sentence to clarify: "We trained all CNN models using Keras (version 3.10.0) application programming interface for TensorFlow (version 2.19.0) in Python (version 3.11.5)."

L248-249 did you use max pooling?

Yes, we have now clarified this by adding "via max pooling" here: "Hence, in two convolutional blocks, the 513 x 513 panchromatic data were downsampled via max pooling to 257 x 257 (62 cm resolution) and then to 129 x 129 (1.24 m resolution), matching the resolution of the multispectral and interpolated DEM data (Figure 2)."

L320-339 please revise; maybe structure like cross-entropy first, then Gerrity

We have revised the topic sentence for clarity: "As part of the hyperparameter experiment, we trained the models with two loss functions: cross entropy and the Gerrity score." Does this help?

L368 please consider rewording the section header; maybe Explainable AI

We split the difference and revised the header to "Explainable AI: Permutation Tests". As permutation tests are only one of many XAI methods, we left the more specific title for clarity.

L386 to me "forever" sounds odd?

Okay, we have rephrased to “for the duration of the test”. We rephrased all instances of “forever” hyperboles in the permutation test description.

L441-444 please consider moving to discussion

We would prefer not to move this to the discussion. The sentences in question provide a contextual overview/synthesis of the results, but do not go so far as to draw conclusions about the overall modeling work—the effectiveness of the method or its potential applications. The only way to connect the confusion matrix to the XAI results, meaningfully, is to consider the objects being identified/confused. Given that the reader is not likely to be familiar with the species or system in question, it makes sense to provide the context along with the synthesis summary of the results rather than leave all context for the discussion.

L450 (Fig 6) could be part of the supplements?

We disagree—we believe it is important to show an example of the data being used to train the CNNs. Figure 6 does triple-duty in showing examples of the image chips described in the methods, the 10 channels, and a contextual example of the model performing well.

L453-455 please consider moving to discussion

Please see our above response on describing results in synthesis.

L458-460 please consider moving to discussion

Please see our above response on describing results in synthesis.

L465 (Fig 7) could be part of the supplements?

We disagree—as we showed a best hit, we should also show an example of a worst miss so that the reader can get a sense for cases where the model is performing poorly. There are many other examples provided in the supplements. A scan of these anecdotes, in combination with the permutation test results, can help with understanding how the CNN is using the data.

L469 this is super interesting, that elevation was the most important predictor.

We agree! We discuss this later in the discussion.

L471-475 please consider moving to discussion

Please see our above response on describing results in synthesis.

L561-564 could you give a reference for that?

We have clarified that these are our direct observations of the study area as opposed to references to general literature. However, in each case we added citations supporting that the observations hold for the species at treeline more generally.

"The species of willow classified here (*Salix glauca* L., *Salix brachycarpa* Nutt., and hybrids) are most abundant near creeks in the Longs Peak study area, with aspen not much farther away (Cooper, 1908). This may generalize when the model is applied at a broader geographic scale; willow and aspen tend to be found in regions with more moisture (Baker, 1989; Coladonato, 1993; Uchytil, 1992; Howard, 1996). Engelmann spruce and subalpine fir tend to be found further from creeks, but in topographic depressions with late-lying snowpack (Burns and Honkala, 1990; Hessl and Baker, 1997; Gill and Campbell and Karlinsey, 2015). Glandular birch and limber pine are more often found on slopes, and limber pine occupies windswept ridges with early snowmelt (Mccune, 1988; Ulrich et al., 2023; Steele, 1990; Cooper, 1908)."

L570-571 could you give a reference for that?

We have added references – please see above.

L589-593 please give a reference for Clever Hans problem

We have added references: "The importance of the elevation data in species classifications, despite its low original resolution of 10 m, suggests that these models could have learned idiosyncratic topographic patterns at the two study sites (i.e., the "Clever Hans" problem, where machine learning models pick up on signals in the data that are sometimes collinear with more meaningful signals, much like the popular historical horse, Clever Hans, who could respond to emotional anticipation in humans but could not, in fact, count or do math) (Lapuschkin et al., 2019; Pfungst, 1911)."

Technical corrections

L20 Please consider the naming convention of scientific species names wrt the authority for the binomial name according to the journal guidelines.

We have added naming authorities for all Latin names.

L68 please consider rewording, the "do" seems off

We have deleted "do".

L95-96 please consider rewording

We have rephrased "often comes at the cost of" to "has".

L 129 blue (450-510nm) lacks a space between number and dimension

We thank the reviewer for catching this.

L343 … and six loss functions …

We thank the reviewer for catching this.

L354 dots represent …?

The caption already says “the black ellipses represent the 3rd through 95th models”. We fixed the spelling typo and added “(dots)” for additional clarity given the orientation of the ellipsis could cause confusion.

L429 it should refer to Tab1

We thank the reviewer for catching this.

L431 reference to Eq. B2 can be omitted

We thank the reviewer for catching superfluous detail.

L511 the term naïve has not yet been introduced up to now, maybe omit

Okay, we have done so.

L555 (Sindewald, Tomback and Neumeyer, 2020)

We thank the reviewer for catching this.

L596 last sentence can be omitted

We thank the reviewer and have done this.

L636 suggestion: We propose an NDVI filter…

Based on the reviewer’s earlier comments requesting the results of this future work, we have deleted this portion of the manuscript to avoid confusion.

L647 (caption Fig A2): reference to Figure A1

We thank the reviewer for catching this.

Citation: https://doi.org/10.5194/egusphere-2025-970-RC1

Reviewer 2 Comments

“General Comments

Your work is an ambitious attempt to tackle the notoriously challenging task of satellite-based tree species classification in an environment particularly well suited to remote sensing applications: alpine treeline systems. You implement a thoughtfully structured methodological framework based on a pixel-based convolutional neural network (CNN), combining 8-band WorldView-3 multispectral data with the higher-resolution panchromatic band and a resampled digital elevation model (DEM) as input.

The study addresses a highly relevant research problem with a novel and exploratory design. Appropriately, the focus lies on identifying workable class compositions and optimal hyperparameter settings. It is clear you are aware of the challenges, as reflected in your use of less common but more appropriate evaluation metrics. For the most part, you discuss limitations openly and express a commendable willingness to improve your framework over time.

The manuscript also sets a high standard for academic writing, particularly for a heterogeneous audience. The language is clear and accessible (e.g., consistent use of active voice), and technical terms are well explained. Notably, you go to considerable lengths to make the CNN methodology comprehensible to non-experts. In a field increasingly saturated with deep learning studies lacking such clarity, this effort stands out as especially praiseworthy."

Thank you for your favorable review! We are gratified that our efforts to make the work clearly understandable to an audience with mixed familiarity are appreciated, especially by someone clearly familiar with ML methods themselves.

"That said, I would like to raise several methodological concerns that I believe require major revisions:

1) Sample size and independence

The modeling workflow relies on a highly imbalanced dataset with only 615 labeled samples from two sites. While you extract 5,631 pixels from these polygons, pixel-level data are strongly autocorrelated—both spectrally and spatially. This violates the assumption of independent and identically distributed (i.i.d.) samples, which is foundational to most supervised learning methods, especially deep learning."

We understand that pixel-level data are strongly autocorrelated. This is one reason that, in Table 2, we report both the number of polygons and number of pixels for each species. We understand that, due to autocorrelation, even if our nominal sample size is 5631, our effective sample size is closer to the number of polygons, namely 615.

We structured the training data around pixels rather than polygons, because this is how CNNs work. CNNs are pixel-based models, rather than object-based models – so it made more sense to have the CNN predict the species for every pixel in a polygon, rather than have the CNN predict the species for the whole polygon at once.

Nonetheless, we are aware of potential problems introduced by autocorrelation. The biggest potential problem is contamination (lack of independence) between the training and validation datasets. This is why we separated the training and validation datasets across polygons (*i.e.,* entire individuals or entire trees), rather than across pixels. In other words, if one pixel in polygon A goes into the training data, all other pixels in polygon A must go into the training data. Similarly, if one pixel in polygon B goes into the validation data, all other pixels in polygon B must go into the validation data. This ensures that the training and validation data contain distinct individuals (*i.e.,* distinct trees), so that the validation data can provide an independent assessment of model performance, based on trees that were *not* used in the training process.

We appreciate your concern about the relationship between model size and sample size. For example, a heuristic from classical statistics is "30x parameters," *i.e.,* that the ratio of num_training_samples to num_model_parameters should be at least 30. However, this heuristic is based on classical linear regression, where the model parameters are estimated independently without strong regularization or weight-sharing. In classical linear regression, each model parameter represents one degree of freedom. However, in modern deep learning, the raw parameter count does not directly reflect the model's effective degrees of freedom. Convolutional architectures (like our own) have weight-sharing, meaning that they reuse parameters across spatial dimensions; they include hierarchical feature-learning (*e.g.,* learning at different spatial resolutions via pooling), which reduces the need to estimate parameters independently; and they include strong regularization (in our case, data augmentation, dropout, L_2, and early stopping), which further constrain effective model capacity.

Furthermore, empirical evidence demonstrates that deep-learning models routinely achieve strong generalization performance with sample-to-parameter ratios far below 30 (*e.g.,* Krizhevsky *et al.* 2012; Esteva *et al.* 2017; Raghu *et al.* 2019), including in meteorological applications (*e.g.,* Dueben and Bauer 2018; Chantry *et al.* 2021). Following common practice in machine/deep learning, our model's generalization ability is validated through strict out-of-bag splits (where pixels from one individual tree are contained either entirely in the training set or entirely in the validation set).

Finally, as you mentioned, we would agree that 615 samples is quite a large sample size given the challenges of work at treeline. While the imagery expense limited the geographic

extent of our study, it is adequate for methods development given the careful selection of locations representative of wider treeline in the region. Even given an image of limited size (25 km^2) and trail access, the terrain and effort required to commute to the sites limits the data that can be collected over five years. With further funding and time, we will be able to test these methods on a broader geographic scale. We do discuss these limitations in the discussion.

"Although you apply data augmentation and strong regularization, I remain unconvinced that these measures adequately compensate for the lack of truly independent training examples. This issue becomes even more pronounced in your fewer-class models."

We appreciate your concern. Again, data augmentation increases the nominal sample size much more than the effective sample size. In total, although we turn 615 polygons into 8 x 5631 = 45 048 augmented pixel-based data samples, we are aware that the effective sample size is much closer to 615 than to 45 048. The primary literature on data augmentation also notes that data augmentation does not increase the effective sample size as much as adding new independent samples (*e.g.,* Shorten and Khoshgoftaar 2019).

However, extensive empirical and theoretical work has shown that well designed data augmentation improves model robustness/generalization by exposing the model to plausible input variations, thereby reducing overfitting to spurious features or noise (*e.g.,* Krizhevsky *et al.* 2012; Perez and Wang 2017). Also, for this study we tuned the hyperparameters for data augmentation to optimize model performance on independent, unaugmented validation data. Namely, we performed a grid search over noise level = {0.1, 0.2, 0.3, 0.4, 0.5} x number of augmentations = {2, 3, 4, 5, 6, 7, 8, 9, 10}. We determined that the optimal hyperparameter settings are 0.2 and 8, the values reported in the manuscript. In this experiment, performance on the unaugmented validation data was more sensitive to the noise level than to the number of augmentations. With a smaller noise level (0.1), performance on the validation data deteriorated, suggesting that 0.1 is not quite enough noise to prevent the model from overfitting to spurious features. With a larger noise level (0.3-0.5), performance on the validation data again deteriorated, suggesting that 0.3-0.5 is too much noise, overwhelming useful patterns in the WV3 data. Note that, whenever we do data augmentation, Gaussian noise is applied to the normalized predictors (in z-score units) rather than the unnormalized predictors (in physical units). Thus, a Gaussian noise level of 0.2 has an equivalent effect on all predictor variables.

While the effective sample size remains limited, our data-augmentation strategy contributes meaningful diversity to the training data that complements regularization techniques. Thus, we believe that the combination of DA and regularization enhances model robustness/performance in a way not achievable by regularization alone.

"The reliance on a CNN under these constraints appears questionable, and may lead to overfitting to spatial artifacts rather than learning generalizable species-level patterns. In small-data scenarios, classical machine learning methods (e.g., random forests) are often more robust. Even if you preferred CNNs to integrate the panchromatic band, methods like pansharpening could have made simpler models feasible."

As mentioned above, both data augmentation and regularization mitigate the risk of overfitting to spatial artifacts. Also, our data-splitting strategy (splitting training and validation data across individuals, *i.e.,* whole trees) ensures that our model assessment is based on independent trees, which were not used in the training process.

We acknowledge that in small-data scenarios, classical ML methods like random forests are sometimes more robust, especially when the features are well crafted and low-dimensional. However, in our context, the high-dimensional spatial structure of the panchromatic and multispectral bands is critical for capturing spatial patterns. CNNs are explicitly designed to exploit local spatial correlations and hierarchical features in image data, which enables end-to-end learning of relevant spatial features. By contrast, a classical ML method like random forests would require one of two approaches: [1] flattening the whole image into a feature vector, where each feature is one pixel at one wavelength; or [2] collapsing the image into tabular data, such as summary statistics. Approach #1 would lead to a very high-dimensional feature vector, with very little information being carried in a given feature; approach #2 would lead to the loss of spatial information and would require manual feature-engineering, which may not capture the full complexity of spatial patterns relevant to our task.

Lastly, although pansharpening can create fused high-resolution images, it can also introduce noise and spatial artifacts (Zhang *et al.* 2023). Using the raw multispectral and panchromatic bands directly avoids these artifacts. We discuss pansharpening in more detail in this paragraph, added to the manuscript, now lines 165-174:

""'Pansharpening" is a similarly inappropriate technique for any analysis that relies on data precision; pansharpening artificially improves the spatial resolution of datasets by modelling statistical relationships between coarser-resolution multispectral imagery with finer-resolution panchromatic imagery and interpolating likely values. The result may aid in visual interpretation for humans, but no new real information is being provided to a CNN. Each larger multispectral pixel is a mixture of spectral signatures of objects at the finer resolution; moving to the finer resolution makes assumptions about the relationship between the two datasets that introduce error and distortions into the data (Zhang *et al.* 2023). The coarser-resolution multispectral data may not have enough information for realistic assumptions to be made, and it is likely this approach would inflate the sample

size without adding new information. We chose to use the data in their native resolutions and adjust our CNN architecture to accommodate the several spatial resolutions. This way, we are still allowing the CNN to learn from the higher-resolution panchromatic data without inflating the sample size of our multispectral data."

"I appreciate how difficult field data collection is in remote treeline environments. Nevertheless, I think it is necessary to explicitly discuss the sample size limitations and model suitability in more depth."

We agree. Based on your comment and our above responses, we have added the following passages to the manuscript itself:

"We delineated species polygons over five field seasons (2019-2023) to obtain the largest sample size possible in the time available. Our haphazard selection (vs. strict randomization) of species patches that met our criteria was also intended to enable us to obtain as many samples as possible. Species patches were delineated if they were 1) greater than approximately 1 m in area and 2) purely a single species with no contamination from co-occurring species. Ideally, patches were selected which were separated from other tree and shrub species by alpine tundra or rock, but this was not always possible. If patches were bordered by other species or if two pure species patches adjoined, we paused collection to walk to the opposing side of the patch before resuming collection, creating a straight line to indicate greater caution during the later labelling process to ensure training pixels were selected which did not contain any of the adjacent species."

We have added this sentence to the beginning of section 2.5, which later goes on to discuss the strengths of CNNs as spatial feature detectors: "We chose to use CNNs to leverage their ability to detect features and relationships between features in spatial data—something traditional neural networks and random forest models cannot do. CNNs are explicitly designed to exploit local spatial correlations and hierarchical features in image data, which enables end-to-end learning of relevant spatial features. By contrast, a classical ML method like random forests would require one of two approaches: [1] flattening the whole image into a feature vector, where each feature is one pixel at one wavelength; or [2] collapsing the image into tabular data, such as summary statistics. Approach #1 would lead to a very high-dimensional feature vector, with very little information being carried in a given feature; approach #2 would lead to the loss of spatial information and would require manual feature-engineering, which may not capture the full complexity of spatial patterns relevant to our task."

We have added this paragraph to section 2.5.2 of the methods (formerly 2.5.1):

"It is important to note that data augmentation does not change the true or effective sample size of our dataset, which remains 615 (the number of tree/shrub polygons) (Shorten and Khoshgoftaar, 2019) However, well-designed data augmentation improves model robustness/generalization by exposing the model to plausible input variations, thereby reducing overfitting to spurious features or noise (e.g., Krizhevsky et al. 2012; Perez and Wang 2017). We performed an initial hyperparameter experiment (see section 2.1 of the supplement) for the data augmentation step prior to our larger hyperparameter described in the next section. While the effective sample size remains limited, our data-augmentation strategy contributes meaningful diversity to the training data that complements regularization techniques. The combination of DA and regularization enhances model robustness/performance in a way not achievable by regularization alone."

We have added this sentence regarding the sample size limitations to section 4.3 of the discussion, which already discusses generalizability limitations and uncertainty: "However, these methods are not sufficient to overcome the inherent limitations we faced with a relatively limited dataset (615 individual tree and shrub polygons) in a relatively limited geographic area (1.75 km2 of a 25 km2 image)."

2) Radiance or reflectance

"The manuscript appears to use radiance values rather than converting to surface reflectance. Although Maxar's ACOMP correction was applied, the resulting data remain in radiance units and are not terrain-normalized. In mountainous terrain, topographic illumination effects can introduce substantial radiometric variability that may bias classification. This is especially problematic given your limited training data, which may not allow the model to disentangle illumination artifacts from meaningful spectral differences.

This choice is not well justified, and the manuscript frequently uses "radiance" and "reflectance" interchangeably, including in the supplement. Please clarify exactly what was used and standardize the terminology throughout the text."

We thank the reviewer for catching this and apologize for the confusion. We have corrected the two instances where "reflectance" was mistakenly used in the manuscript and the captions in the supplement ("species reflectance curves").

The data were in units of surface radiance rather than surface reflectance. However, the data had been corrected for atmospheric effects and terrain effects via Maxar's ACOMP correction. (I.e., the data were not in units of top-of-atmosophere radiance.)

(https://pro-docs.maxar.com/en-us/ReleaseNotes/Pro/2023/2023v1_4.htm#:~:text=Atmospheric%20Compensation%20(AComp)%20Maxar',and%20supports%20feature%20extraction%20applications)

The remaining conversion accounts for WV-3 sensor differences in the maximum solar exoatmospheric irradiance detected by the sensor array in any given band and accounts for solar illumination angle at the time of the image collection. Converting to surface reflectance normalizes radiance values by the maximum values detectible by the sensor, effectively removing these sensor artifacts. We agree that, for the purposes of applying the model to other imagery or the data to other systems, the data should be in unitless surface reflectance rather than radiance.

However, we do not think that our models need to be retrained on the normalized values at this time. We examined how the species spectral curves shifted (please see the below examples), and the relative differences between species remained consistent (though the magnitude of the differences declined). This was true for all species comparisons. Because our training and validation data were both in the same image with the same relative differences, we do not believe the CNN model results would change substantively were they to be retrained on data in surface reflectance as opposed to surface radiance.

Please see Figures 1-32 at the end of this PDF, comparing the data pre- and post-conversion (surface radiance vs. surface reflectance).

3) Handling of background and edge effects

"The sample preparation process lacks detail regarding how background pixels and edge effects were handled. You state that all input image tiles were taken fully within species polygons, which implies no edges were included. If so, the model is only exposed to physically contiguous vegetation, and would not learn features of solitary trees or species transitions, which are common at treeline.

Moreover, interior pixels may still include non-vegetation (e.g., soil, rock, shadows), especially given the scale of the imagery and the heterogeneous terrain. Without erosion, masking, or validation of pixel content, the labels may include noise. This is a non-trivial concern in sparse, high-elevation environments. You should either justify this omission or describe your handling of it more thoroughly."

We have added clarification on our patch sampling methodology. We did avoid mixed-species pixels, but if anything, our sampling process included *more* solitary trees and/or single-species patches than multi-species krummholz patches. If multi-species krummholz patches were large enough and were not intermixed (i.e., adjacent single-species patches), we did attempt to collect them. Please see the quoted section above (of added lines 190-199) regarding patch delineation.

It is possible that the model struggles more with species in less common contexts, or in contexts not represented by the dataset. We discuss this with respect to the model's best hits and worse misses in the main manuscript and in section 5 of the supplement.

It is true that the labels may still include noise, but we believe this provides additional information to the model. Limber pine canopies are sparser, so those pixels are likely to contain more gravel and perhaps co-occurring forbs, graminoids, and cushion plants. Engelmann spruce and subalpine fir form true krummholz mats with little to no ground exposure. The "noise" in the spectral reflectance data at the canopy level will be consistent/limited within a species category, will invariably be present at this resolution, and it is therefore important that we capture/represent it.

4) Narrow hyperparameter search space

"Although you describe your hyperparameter tuning as extensive, the search space for critical regularization parameters was surprisingly narrow. Only two dropout values were tested (0.575 and 0.650)—both high—and L2 strengths were confined to a narrow range around 1e-6. While strong regularization may be reasonable given the small dataset, it is unclear why other plausible values were excluded.

Without a broader search, your claim of extensive hyperparameter tuning is weakened, and it becomes unclear whether the final configuration reflects an optimal balance."

Thank you for noticing this. The early work for this study involved two other hyperparameter experiments, which were not discussed in the manuscript, for the sake of brevity. One early HP experiment involved data-augmentation settings (see our response to major comment #1). The other early HP experiment involved regularization settings, in which we experimented with the optimal values for L_2 weight, dropout rate, and number of dense layers. Specifically, we performed a grid search over L_2 weight = {10^{-7}, $10^{-6.5}$, 10^{-6}, $10^{-5.5}$, 10^{-5}}; dropout rate = {0.500, 0.575, 0.650, 0.725, 0.800}; and number of dense layers = {3, 4, 5, 6, 7}. We trained a separate model for all 5 x 5 x 5 = 125 combinations of these three hyperparameters. We made the following conclusions from this early experiment:

- The optimal number of dense layers was 3 or 4, while values of 5-7 led to overfitting. Since 3 and 4 are on the edge of the search space, for the experiment in the manuscript, we decided to experiment more widely, also training models with 1-2 dense layers.
- The optimal dropout rate was 0.575 or 0.650.
- The optimal L_2 weight was $10^{-6.5}$ or 10^{-6}.

We have revised the manuscript with these details (in the Supplement, section S2).

5) Figures and visualizations

"The manuscript omits basic training diagnostics, such as loss and accuracy curves over epochs, which are essential for assessing convergence and potential overfitting."

We have plotted these learning curves; please see section S3 of the supplement. The learning curves show that, despite the models overfitting the training data, they converge for the validation data. Without regularization (namely data augmentation, L_2 regularization, and dropout), the learning curves looked much different. Specifically, for later epochs, all metrics improved on the training data while deteriorating for the validation data – as opposed to converging for both datasets.

"You also do not include visual comparisons of model predictions against ground truth (e.g., classification maps or overlays). Given the effort devoted to supplementary content, these are surprising omissions.

I strongly encourage including at least one visual example of classification results—preferably showing raw input, model output, and ground truth. This would improve both transparency and interpretability of your results."

We have plotted these figures; please see Appendix D of the manuscript.

6) Model reliance on elevation

The discussion identifies elevation as one of the most important predictors. You note the risk of learning idiosyncratic topographic patterns specific to the two study sites. However, you do not address a critical possibility: that the model's (partly) modest improvement over random baselines may be driven primarily by elevation, rather than species-specific spectral features. This warrants closer examination, particularly in light of your generalization goals.

We agree, but beyond the XAI methods we have already used, the best way to test this is to apply the model to a geographically independent area. Given funding and time limitations, this was beyond the scope of the present work. However, while elevation was often the most important predictor, it was not the *only* significantly important predictor. If the spectral predictors did not contribute to the model's improvement over a random baseline, they would not have significantly reduced model performance when permuted. Please see the results of the permutation tests as evidence that the models could not have performed as skillfully with elevation alone.

We furthermore provide methodology that allows researchers to use panchromatic and multispectral imagery, and a digital elevation model, all at different resolutions *at their native resolutions*, without sacrificing information by downsampling higher resolution

imagery to the coarser resolutions before training. These methods, being so much less expensive than alternatives, will be extremely useful if the models generalize to other areas. We look forward to performing this further test at a later date.

7) Generic CNN explanations

Your effort to explain CNN architecture is appreciated and makes the manuscript more accessible. That said, much of the explanation remains generic. The discussion would benefit from workflow-specific illustrations (e.g., how sample image chips are processed across layers), which could make the architecture and its function more concrete.

While we agree we have quite a lot of generic explanatory detail about CNNs (and are glad that it was a worthwhile effort), we do go into detail on the CNN architecture (our specific methods) in section 2.5.2 (formerly 2.5.1) and in Appendix C. We struggled with the correct balance of detail in the manuscript vs. the appendices and supplement. We received a critique from Reviewer 1 that the manuscript is long with possibly too many figures and tables. However, we agree that a workflow-specific illustration(s) would be helpful and have added it as Figure 2 in the main manuscript.

In conclusion, I believe that the manuscript has merit but requires **major revisions** to address the methodological issues above. I offer these points in the spirit of constructive dialogue and welcome disagreement or clarification from more experienced reviewers. As I am still early in my academic career, I do not claim all concerns are definitive—but I believe they raise legitimate and important questions.

We thank you for your detailed and substantive review. The manuscript will certainly be stronger due to your attentiveness. We will add a note to the acknowledgements to this effect.

Specific comments

L81 “These advances in remote sensing technology have led to an exponential increase in species identification studies since 1990 (Fassnacht et al., 2016; Pu, 2021).” – A statement such as this should include more recent references.

We have cited the most recent review papers existing on the topic of remote identification of tree species, as far as we are aware. We have clarified the sentence with “tree species”.

L85 You should clarify what you are referring to – is it classification, semantic segmentation, object detection, or instance segmentation?

Mishra et al. (2018) performed spectral difference segmentation, which is a form of semantic segmentation. They then classified the segmented image with an object-based

approach. We have revised the lines to read, “Mishra et al. (2018) succeeded in achieving 73% overall accuracy in classifying four tree species across a ~ 140 m x 80 m region of the Himalayas using multispectral UAS imagery. This success highlights the potential and effectiveness of high-resolution UAS data for treeline species identification using an object-based classification approach with spectral difference segmentation (a form of semantic segmentation).”

L86 Please elaborate on “previous methods”.

We have revised “previous methods” to “field surveys”.

L91 Putting it this way, makes it sound like HSI and LiDAR are closely related sensor systems. Please consider a differentiation between optical and active sensors.

Based on the reviews of the literature (and specific studies) cited, these two systems (while clearly very different) are commonly used in concert for the specific task of tree identification. We have revised the sentence to include a definition of lidar requested by Reviewer 1 and to clarify the point you have made here.

“Airborne hyperspectral imagery and lidar (or LiDAR—light/laser imaging, detection, and ranging—an active remote sensing platform that uses time lags in reflected laser pulses to measure distances and so map the 3-D surface structure of sensed objects) have also enabled commonly been used in concert for tree species identification (Dalponte and Bruzzone and Gianelle, 2012; Matsuki and Yokoya and Iwasaki, 2015; Liu et al., 2017; Shen and Cao, 2017; Voss and Sugumaran, 2008), in some cases with classification accuracies ranging from 76.5-93.2% (Dalponte and Bruzzone and Gianelle, 2012).”

L96 Can you move this part up to the optical airborne systems? Since your work is based on satellite imagery, your lineup should ideally end there.

Thank you, we have made this change.

L174 This is the only instance where you refer to the WV-3 data with the term “reflectance” (instead of “radiance”).

There was one other, but either way, thank you for catching it!

Table 1 I think you forgot one of the species (BEGL) here.

Thank you for catching this surprising error! We have corrected the table.

L184-191 I appreciate the intent but I think this “preview” is redundant.

We would like to keep the preview based on Reviewer 1’s comments (and requested addition, at that).

L237-238 This approach artificially inflates your dataset without increasing the true variability.

The manuscript now includes a longer discussion on the difference between effective and nominal sample sizes. See our response to major comment #1, which highlights the passages that have been added to the manuscript.

Figure 6 & 7 I find both figures (this also refers to the ones of the same type you added to the supplements) rather redundant. They do not provide any meaningful visualization of your results. And they have a counterintuitive caption that made me look for a part of the figure that showed the class probabilities. Instead, I advise you to zoom in, show the actual image input (panchromatic & a random band), ground truth, classification result, and potentially the error between classification and ground truth. Then readers are more likely to follow your approach.

A main purpose of each figure is to show all inputs (predictors) and outputs (the prediction and the ground truth) for one case. If we subset the inputs (*e.g.,* by showing only the panchromatic band and one random multispectral band), the figure would not serve this purpose. If we zoomed in and subset the spatial domain, again the figure would not serve this purpose, because it would not show everything going into the model. As for the additional components you asked for – namely the ground truth, classification result, and error between the two – these are all shown in the figure title. Since all these values are scalars (there is only one classification result and one true class – and thus one error – for the entire image), we found that stating these values in the title is a more effective display method than trying to overlay them with the imagery (which would also obscure some of the imagery).

We disagree that figures 6 and 7 do not provide meaningful visualization. The figures are showing exactly the actual image input for example validation cases provided to the CNN. The figures are "zoomed in"—they are 160 x 160 m subsets of the imagery, centered on a validation case. Figure 6 does triple-duty in showing examples of the image chips described in the methods (all of the spatial context provided to the CNN for each training sample), the 10 channels, and a contextual example of the model performing well. The best hits and worst misses allow the viewer to get a sense of the kinds of cases the model does well or poorly at classifying; they were revealing in the context of permutation test results that the model was relying on the panchromatic band (higher resolution textural information) and elevation. The model struggled with cases where a pixel was part of a large, mixed-species krummholz patch and where spruce and fir were both found in abundance (example in Figure 7), but did well at predicting limber pine in a region of the study area where limber pine was abundant and growing in characteristic fashion: solitary

rather than in tree islands with other species, and with a low density of other limber pines nearby (Figure 6).

That said, we have made a figure showing the spatial distribution of the labeled species pixels, as well as panels showing the model's predictions and whether those predictions were correct (see Figures 8-13 of our response to your review).

For clarification, we have modified the figure captions. For example, the caption of Figure 6 now reads as follows:

"Figure 6: Example of a "best hit" classification for the six-class model. For the given data sample, this figure shows all inputs to the model (paneled images) and all outputs from the model (the true class and predicted class in the title). Specifically, the true class is PIFL, which the model correctly predicts with 100% probability. This example (image chip or patch) is from the Battle Mountain study site. All eight multispectral bands, the panchromatic band, and the DEM are shown. The red star in the center of each image patch is the pixel being classified. Units of radiance are W m^{-2} sr^{-1} μm^{-1}."

L587 Since the main methodological concern is whether the small sample size demands lighter machine learning models, you should specify to "deep learning" here.

We appreciate that our sample size is limited, but we do not think it is small enough that it demands a lighter machine learning model (see response to General Comment #1). Furthermore, while deep learning models are prone to overfitting, they are not the only form of machine learning model that tends to overfit data. For example, a single decision tree frequently overfits the data, which is why random forests ensembles of decision trees were developed; each decision tree is biased in some fashion, but collectively the forest of trees may offset their biases and so help to overcome the issue of overfitting.

Technical corrections

L56 delete "important"

L320, 323 imbalanced

We have made these corrections; thank you for catching these errors.

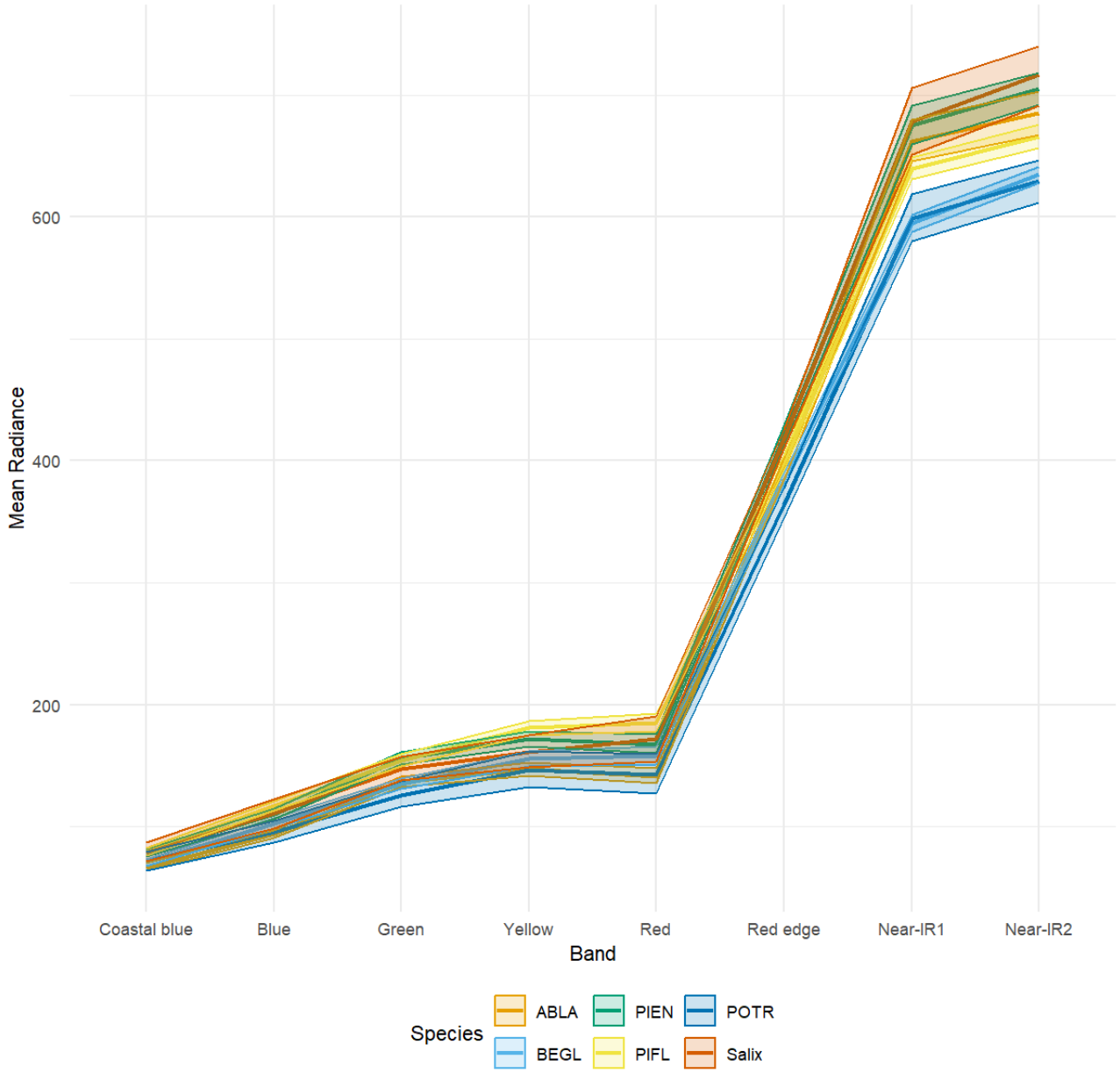

Figure 1. Surface reflectance for all species

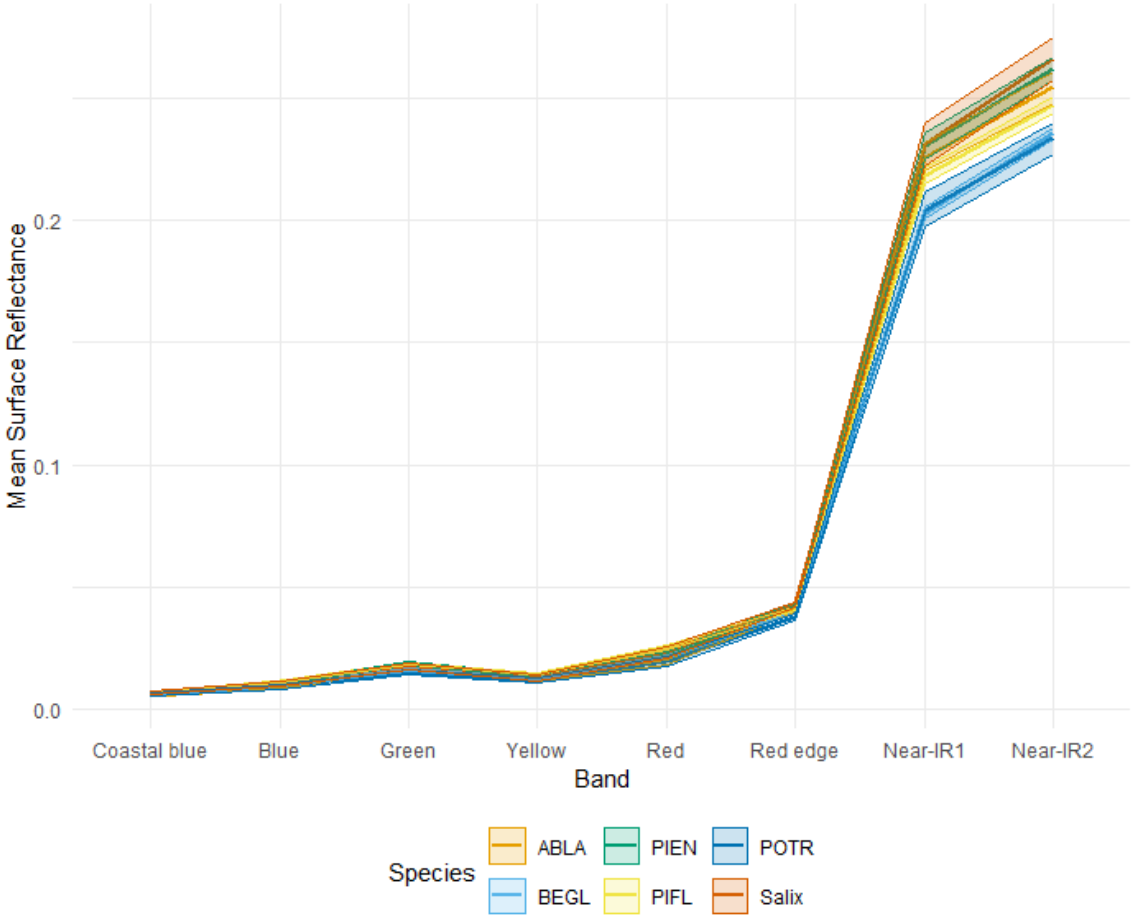

Figure 2. Surface radiance for all species.

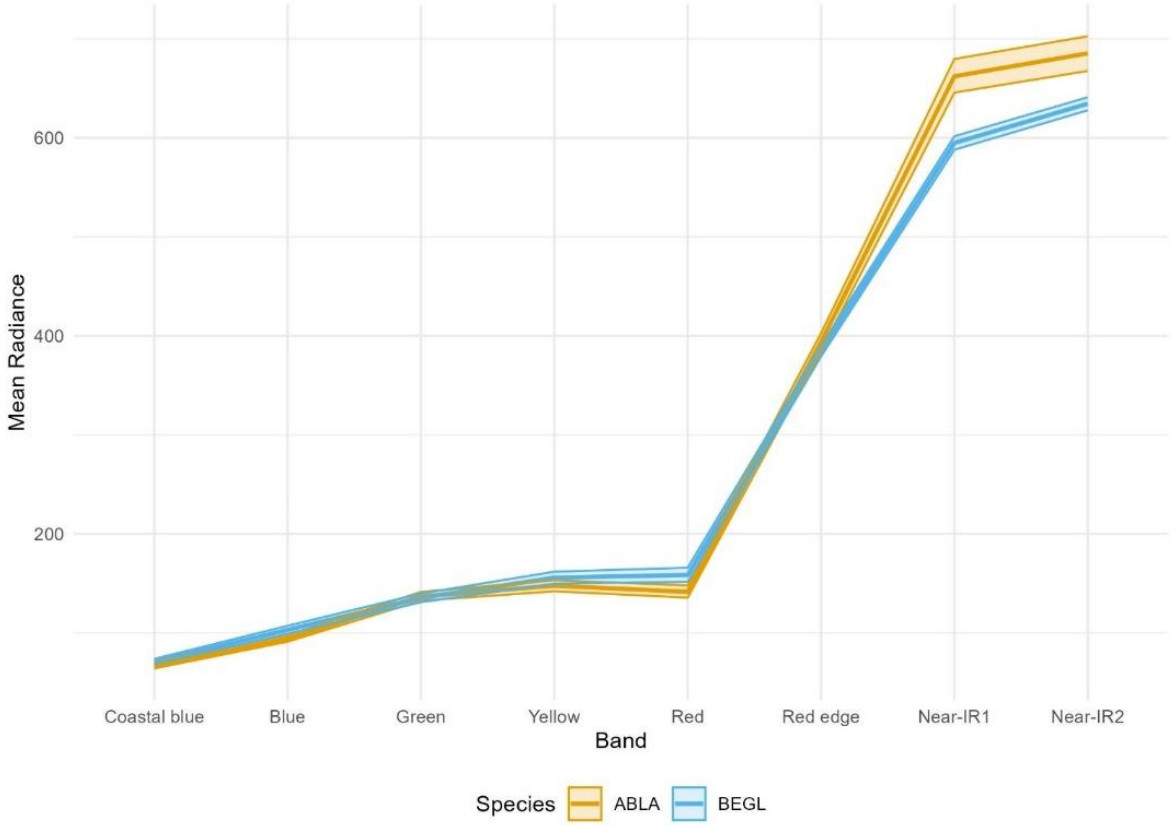

Figure 3. Surface radiance for ABLA vs. BEGL.

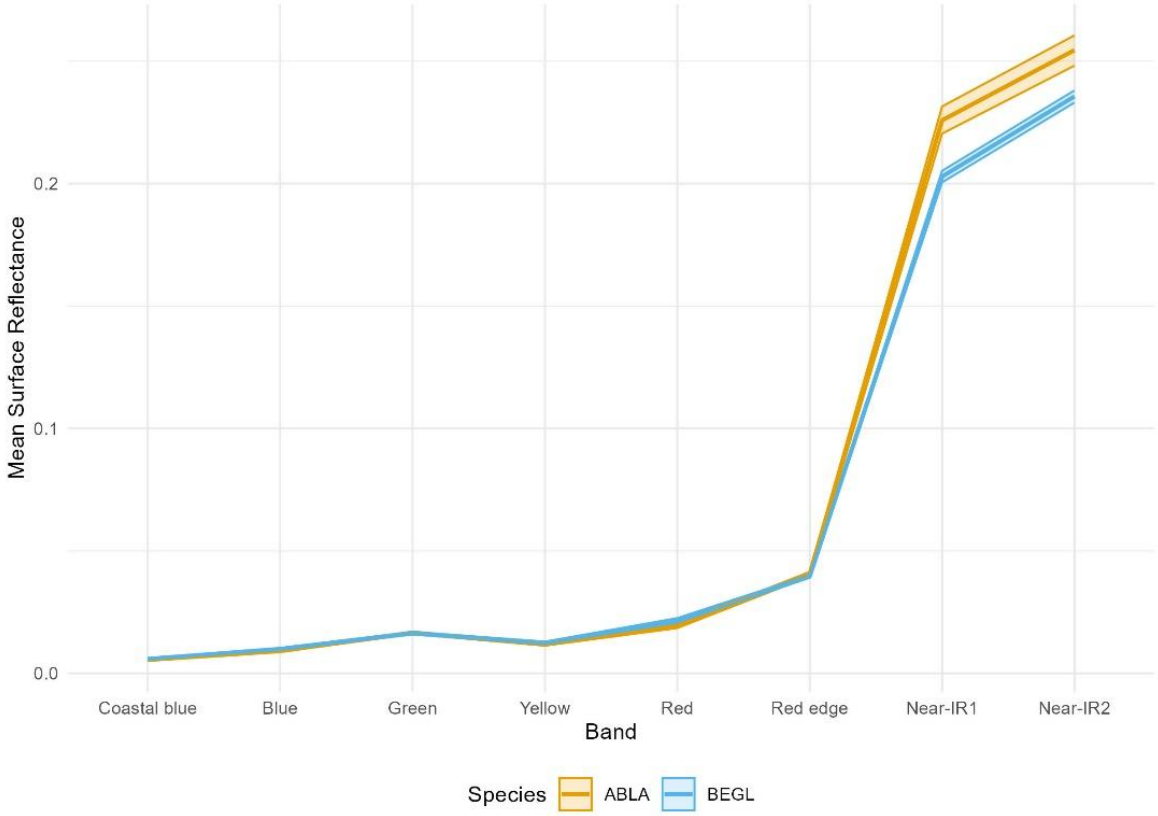

Figure 4. Surface reflectance for ABLA vs. BEGL.

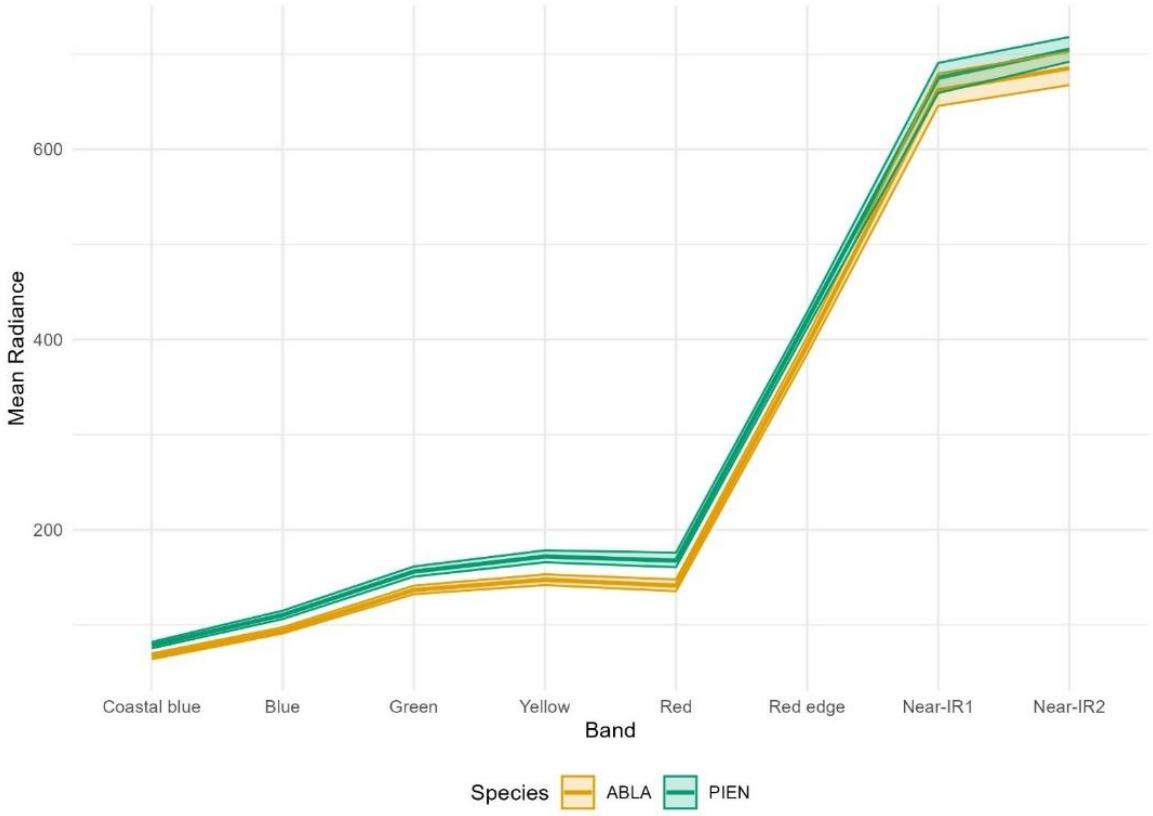

Figure 5. Surface radiance for ABLA vs. PIEN.

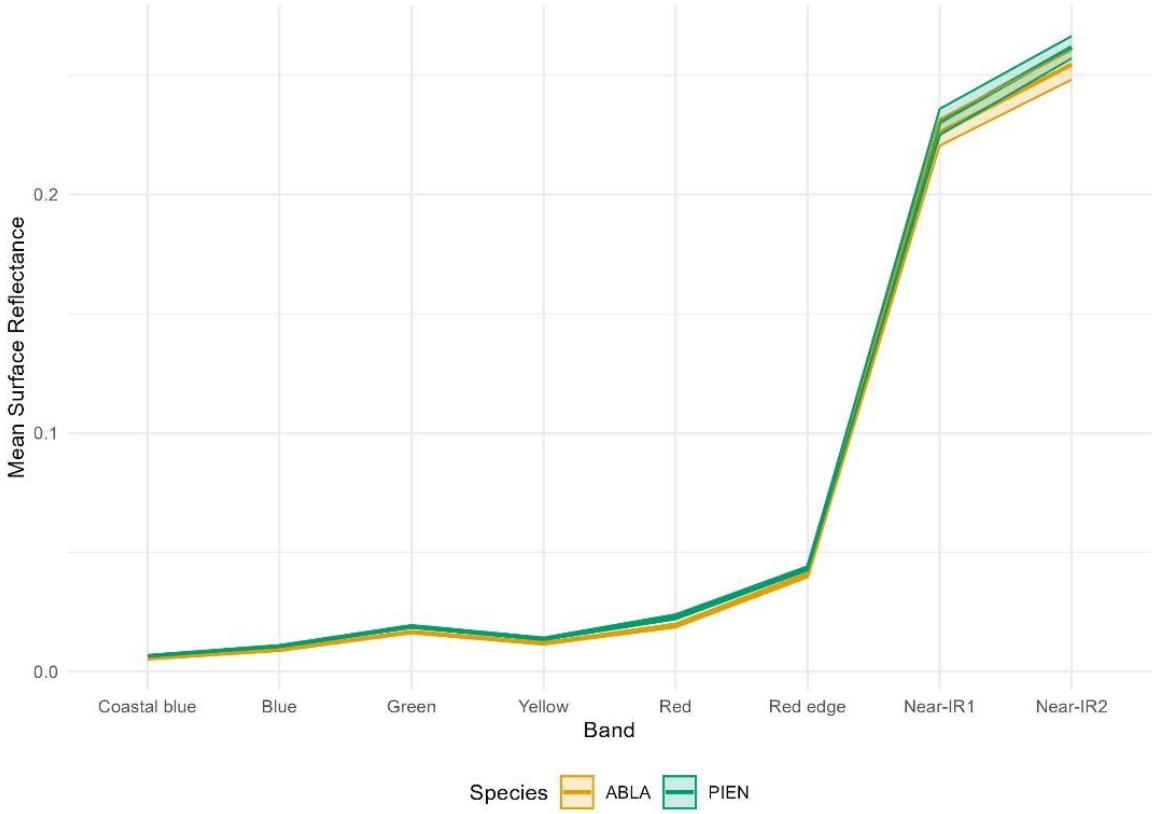

Figure 6. Surface reflectance for ABLA vs. PIEN.

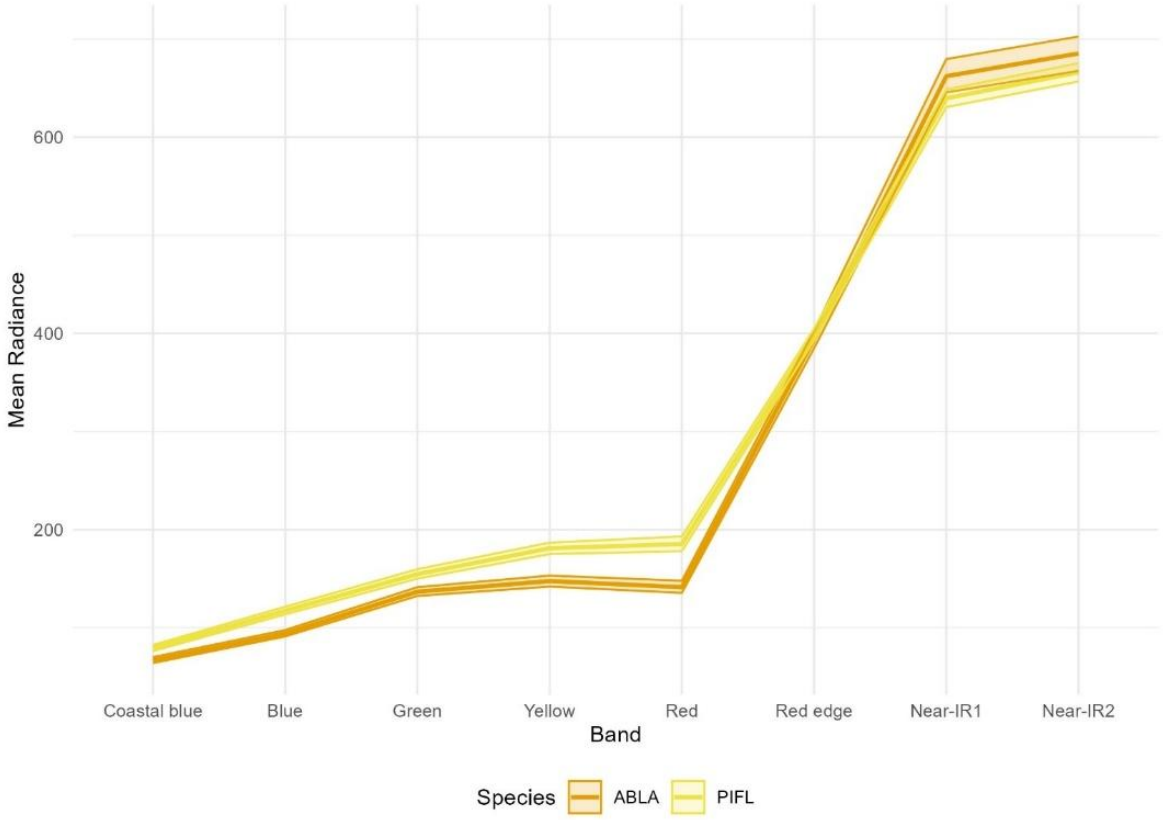

Figure 7. Surface radiance for ABLA vs. PIFL.

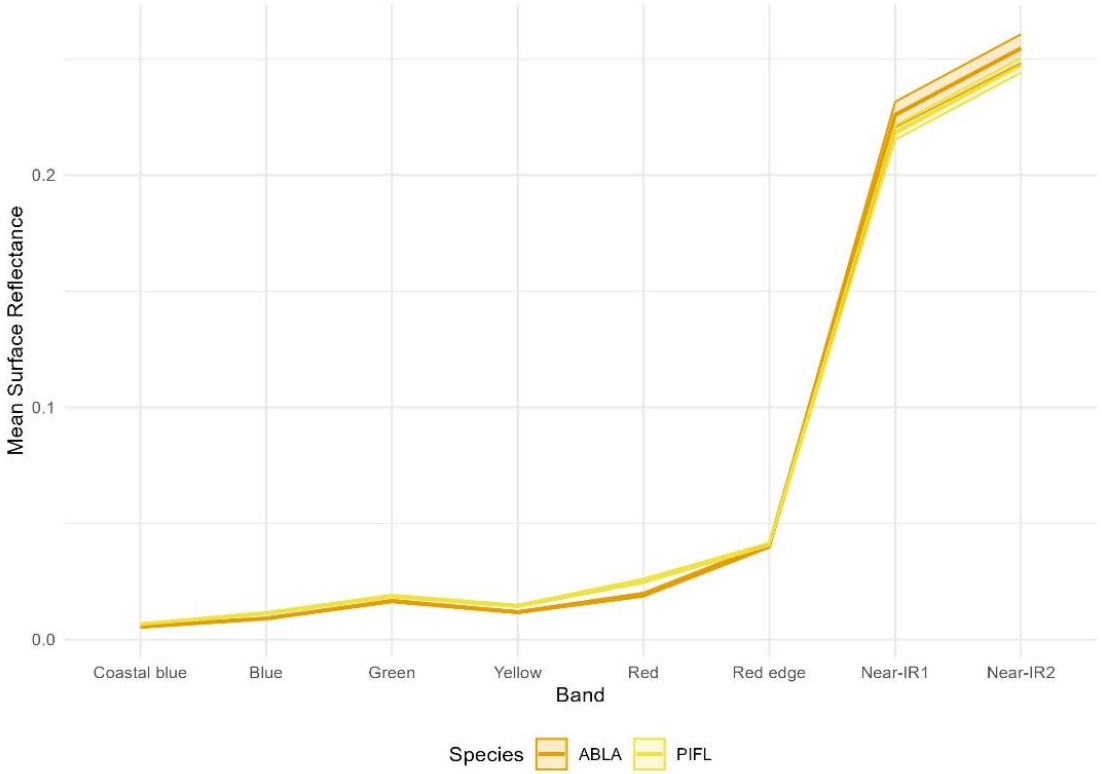

Figure 8. Surface reflectance for ABLA vs. PIFL.

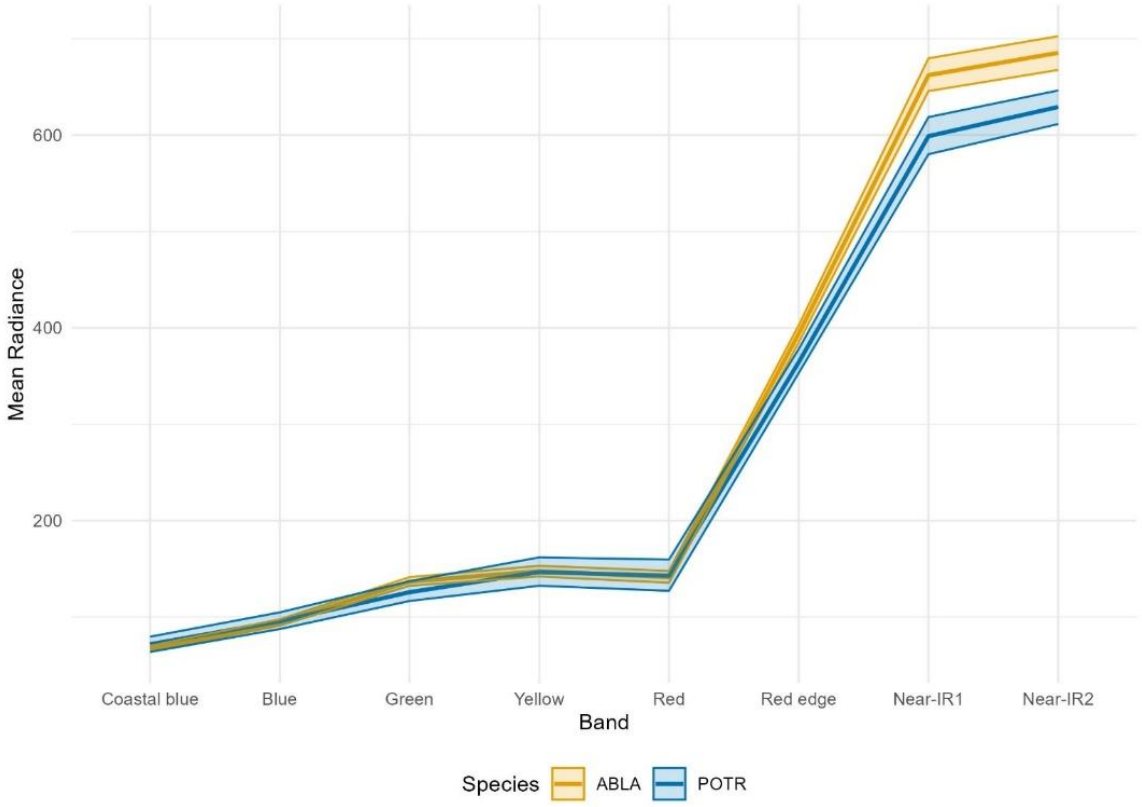

Figure 9. Surface radiance for ABLA vs. POTR.

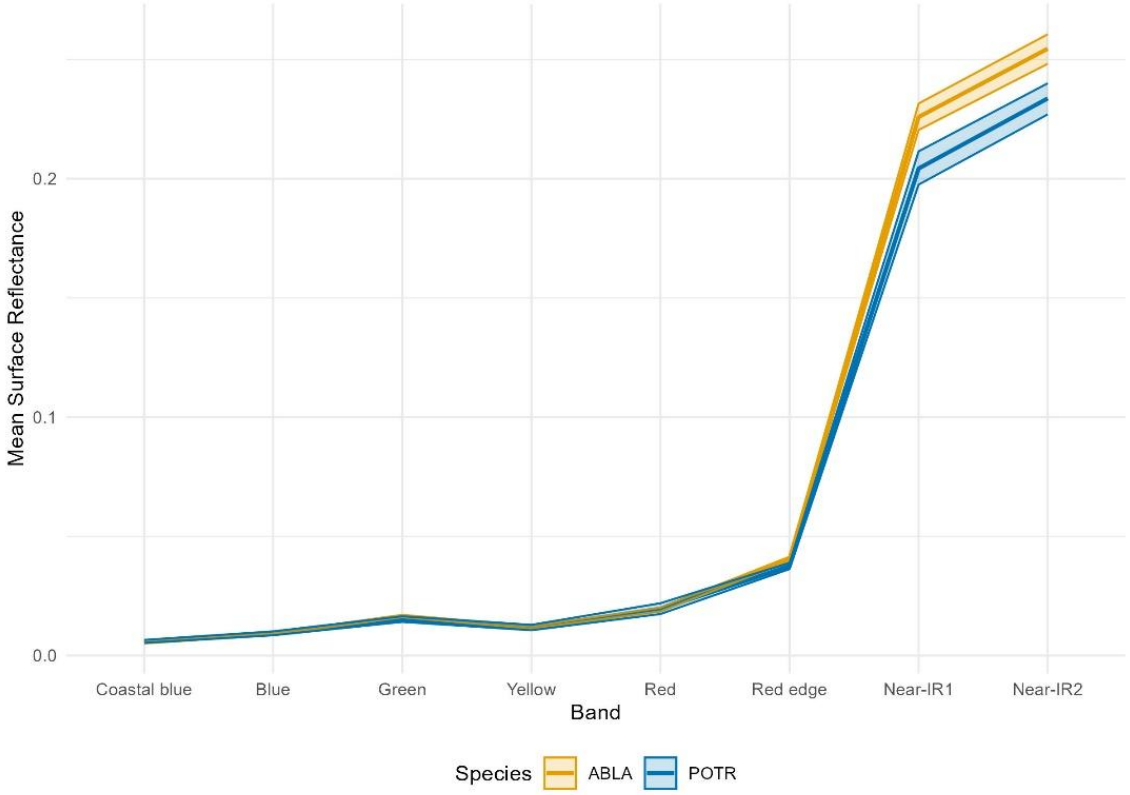

Figure 10. Surface reflectance for ABLA vs. POTR.

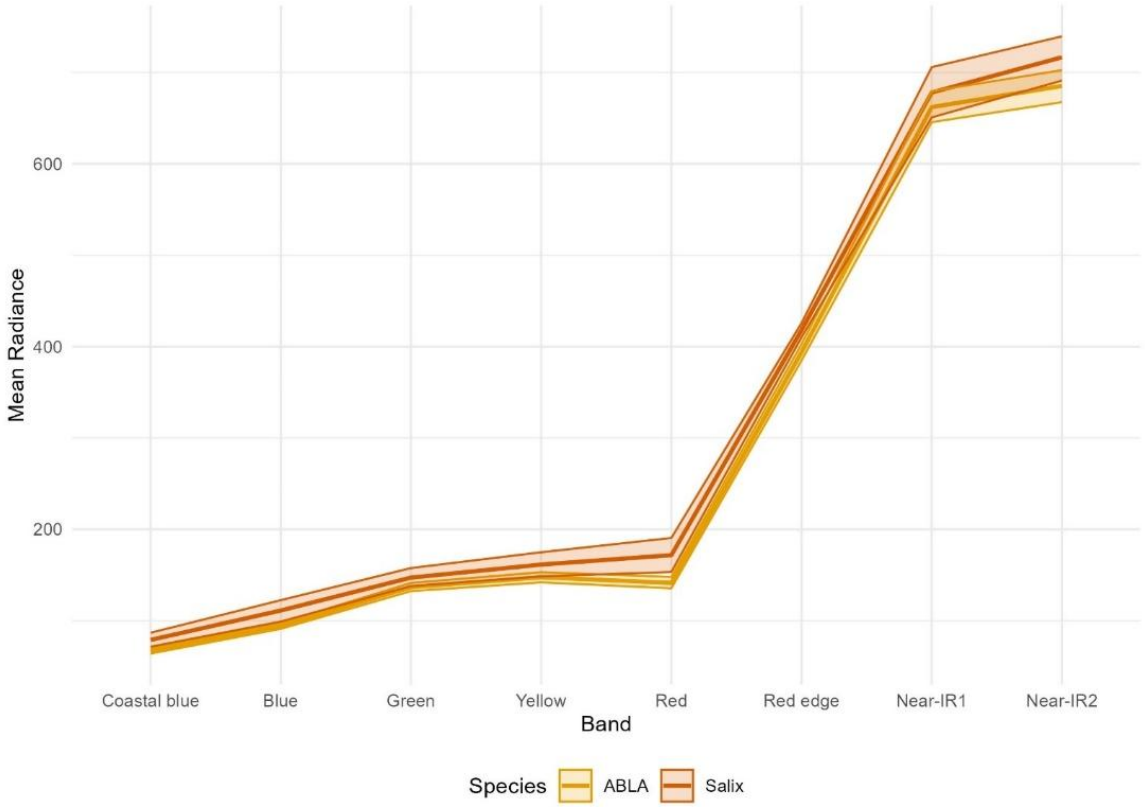

Figure 11. Surface radiance for ABLA vs. Salix.

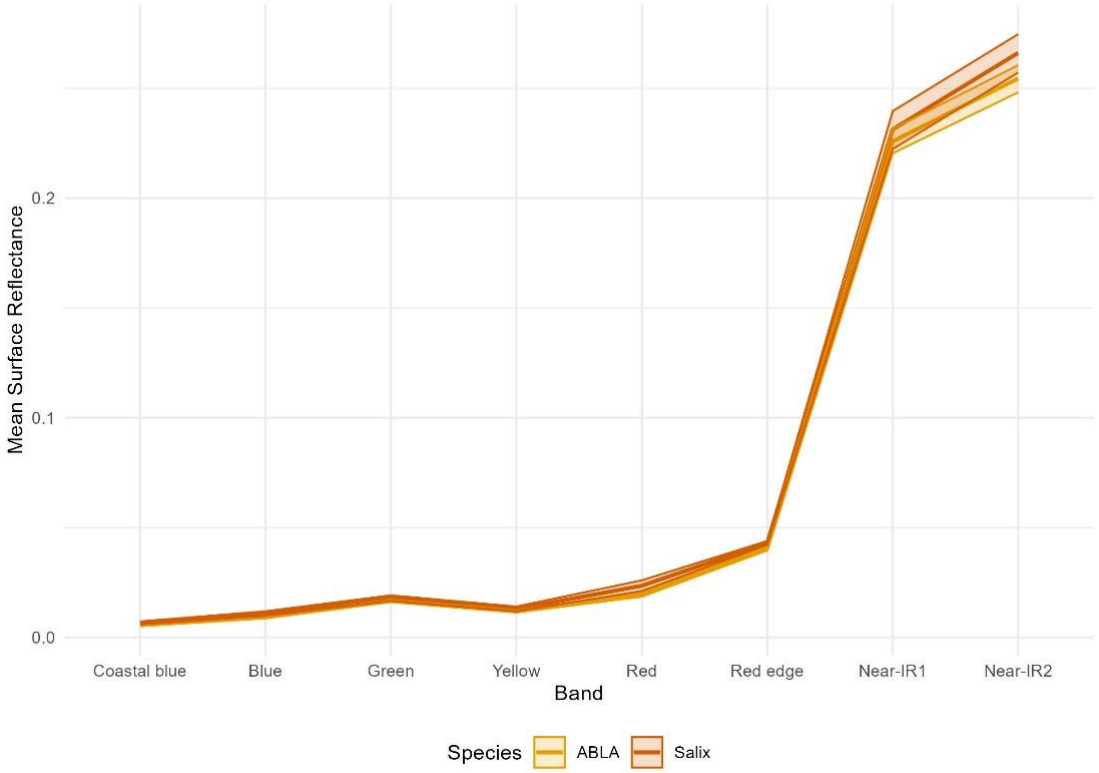

Figure 12. Surface reflectance for ABLA vs. Salix.

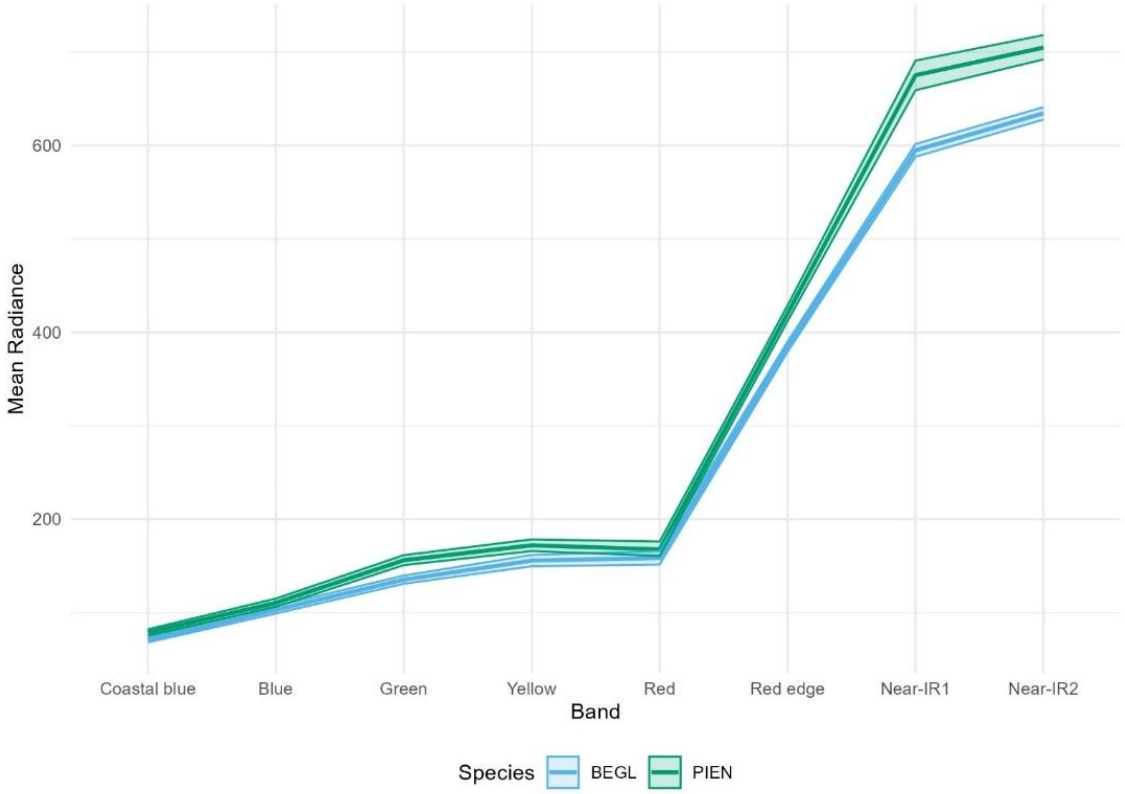

Figure 13. Surface radiance for BEGL vs. PIEN.

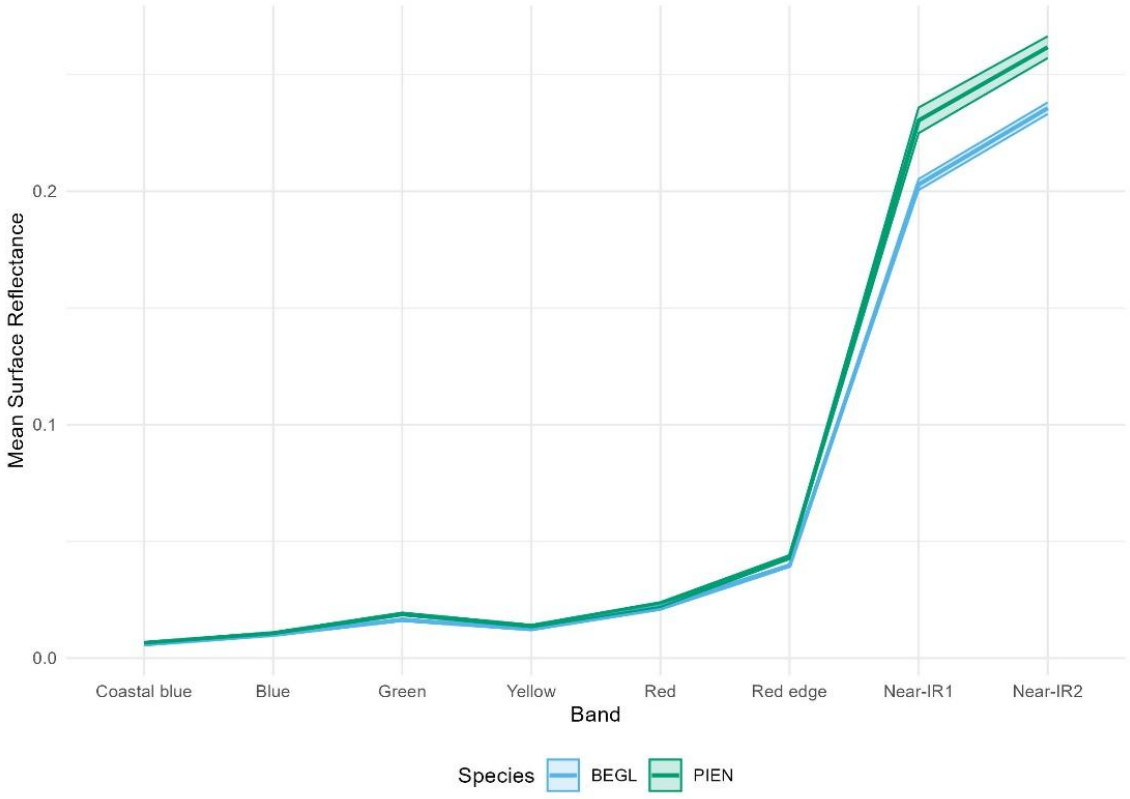

Figure 14. Surface reflectance for BEGL vs. PIEN.

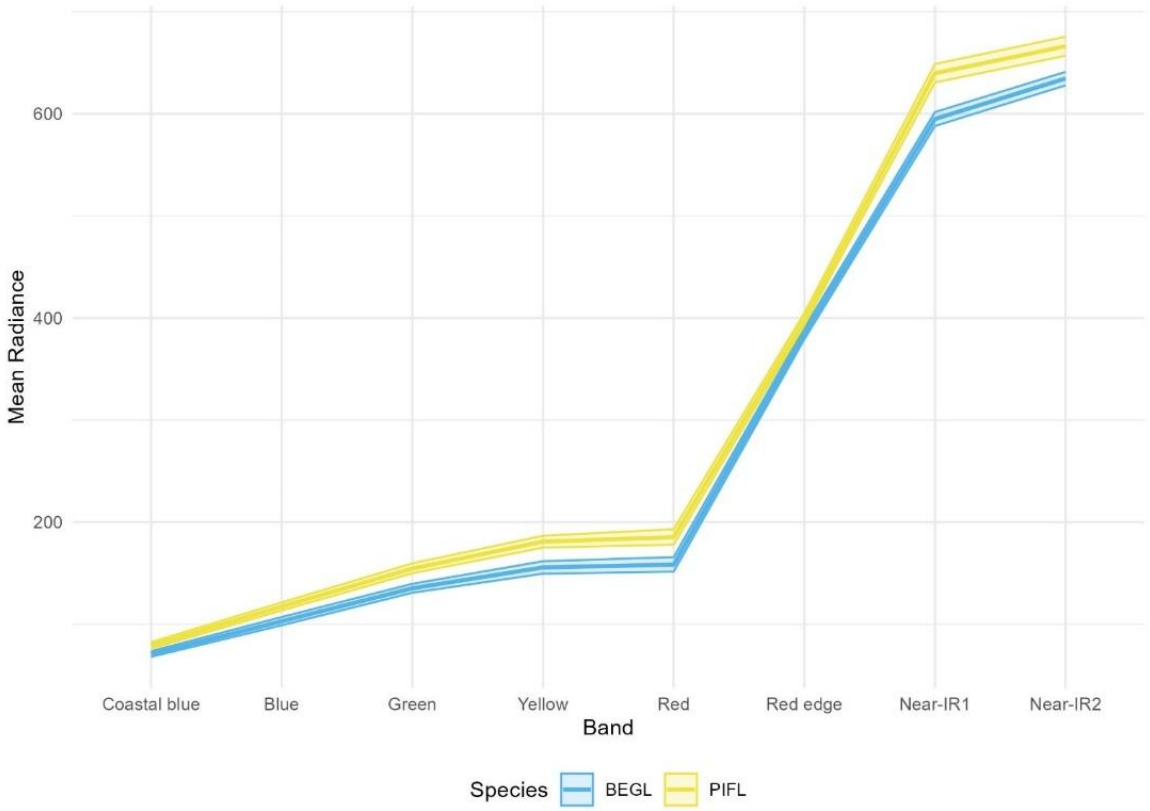

Figure 15. Surface radiance for BEGL vs. PIFL.

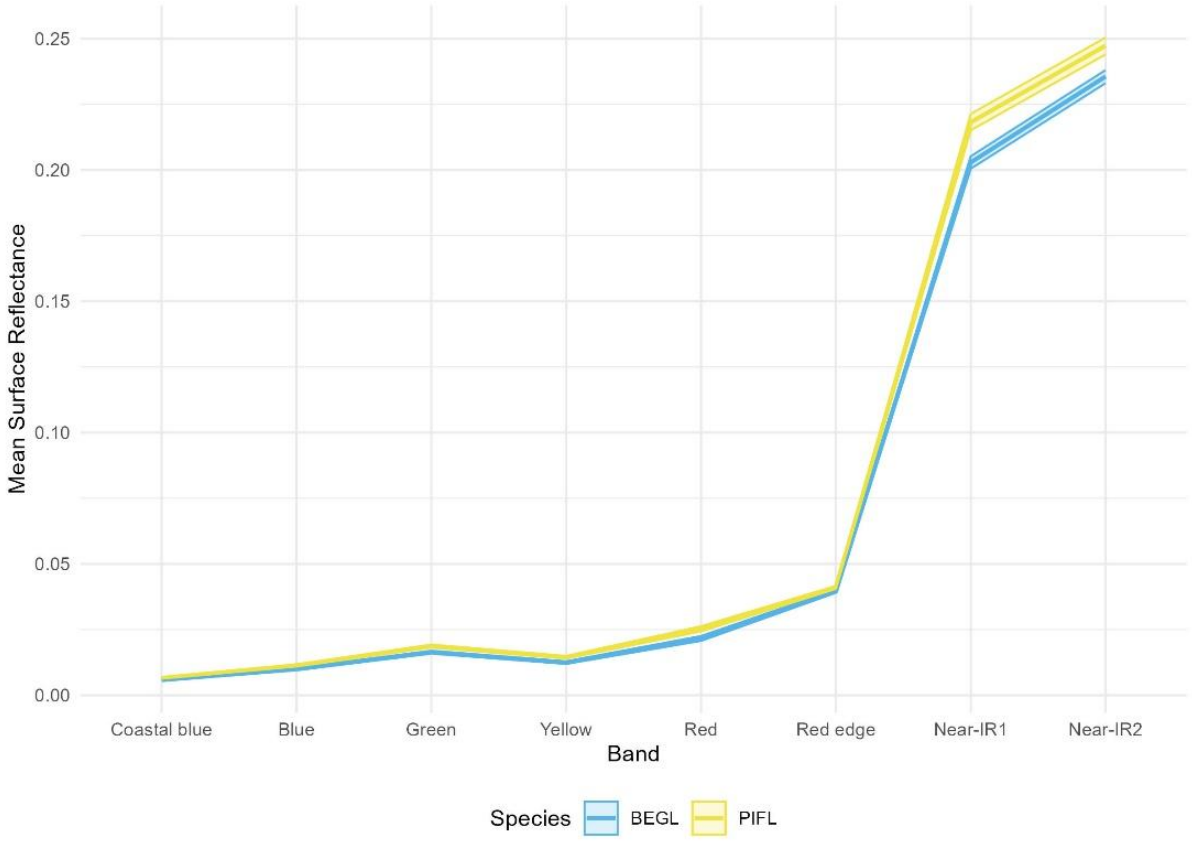

Figure 16. Surface reflectance for BEGL vs. PIFL.

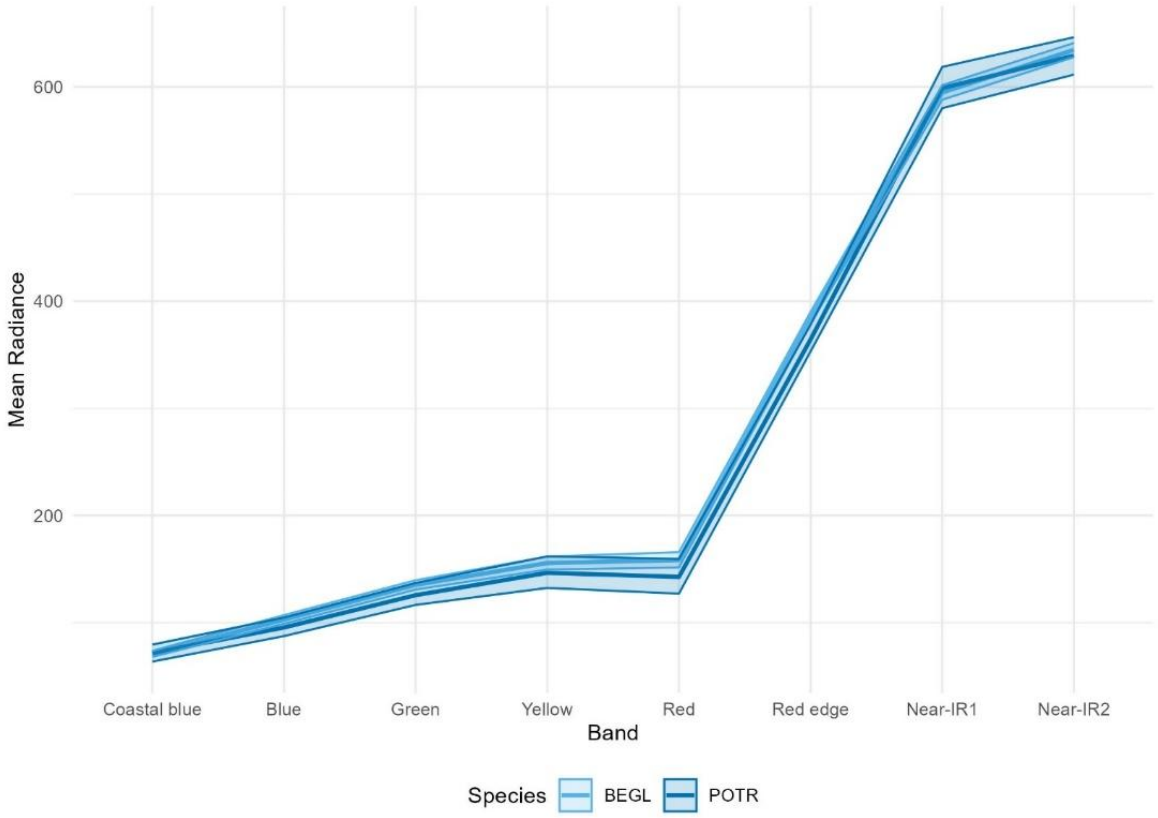

Figure 17. Surface radiance for BEGL vs. POTR.

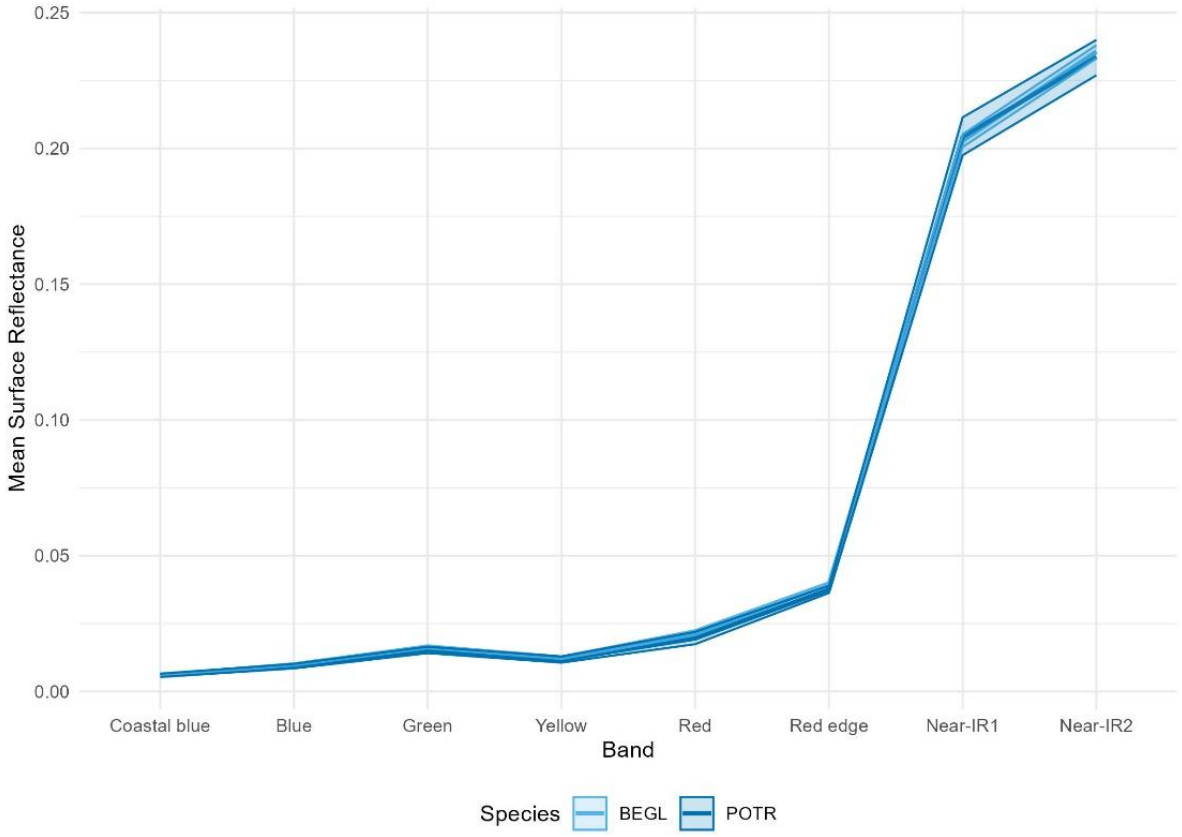

Figure 18. Surface reflectance for BEGL vs. POTR.

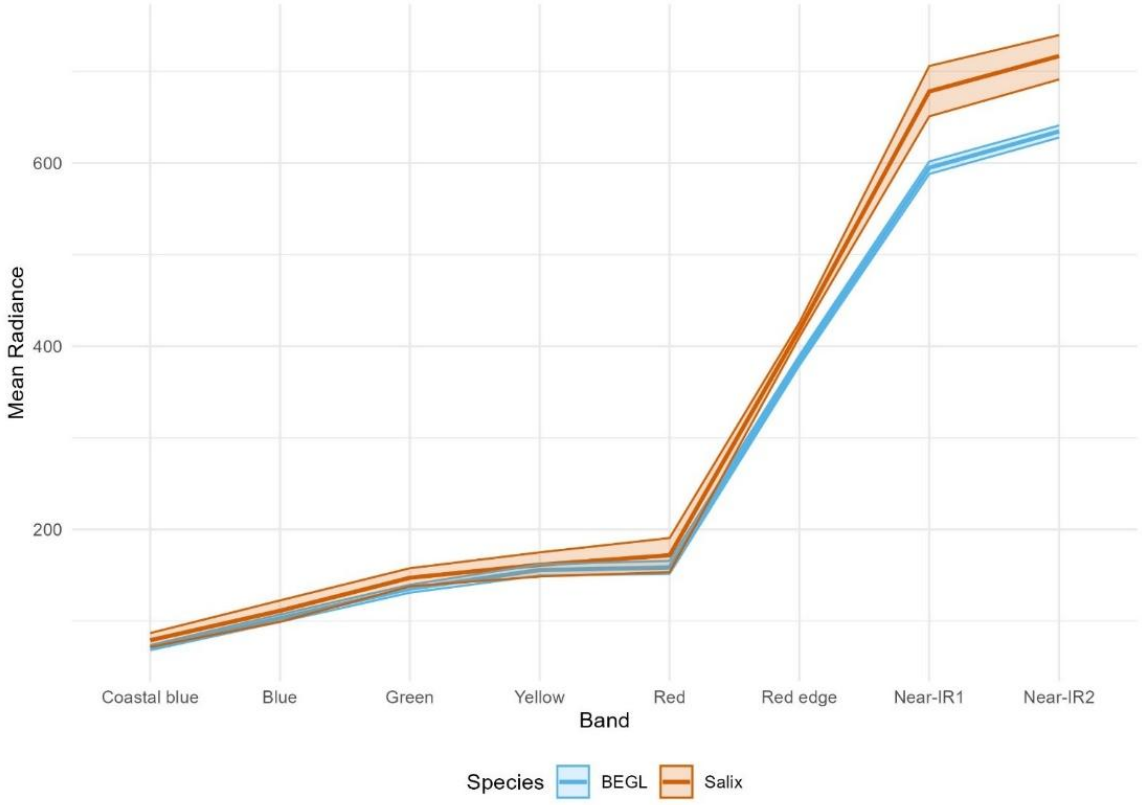

Figure 19. Surface radiance for BEGL vs. Salix.

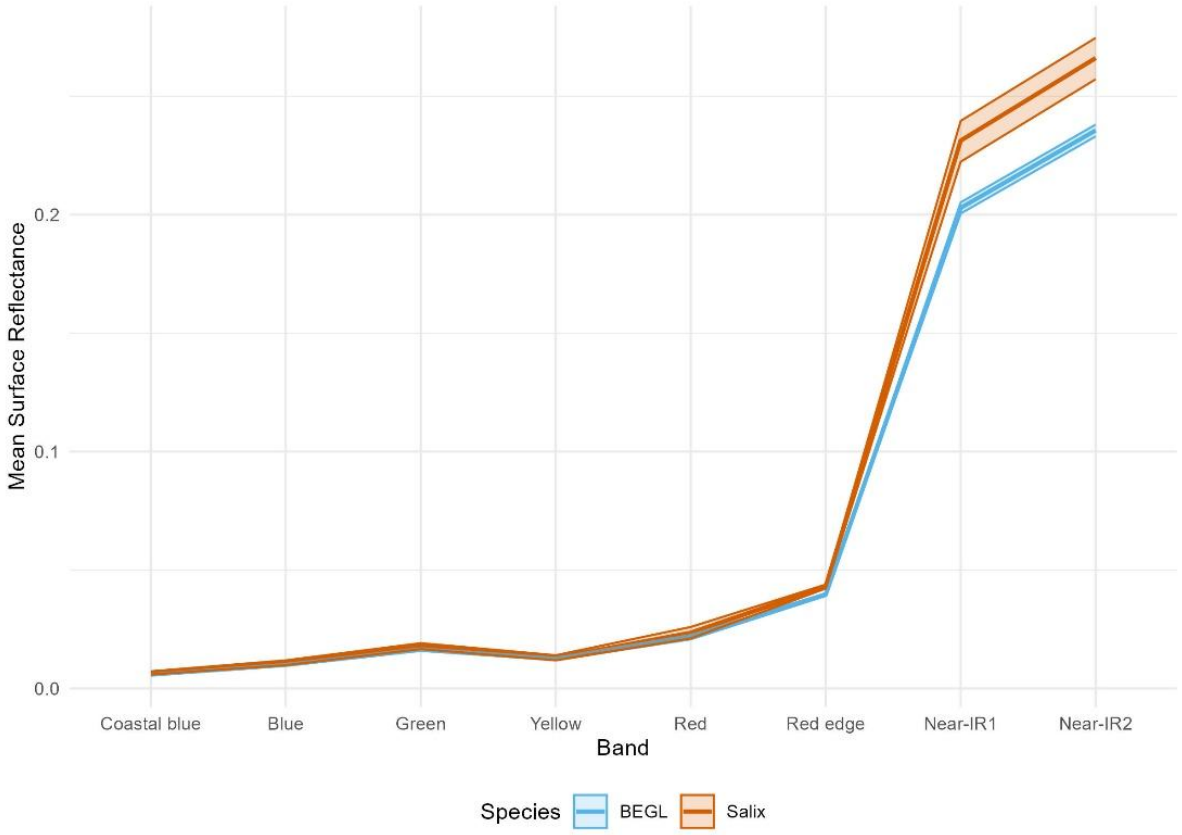

Figure 20. Surface reflectance for BEGL vs. Salix.

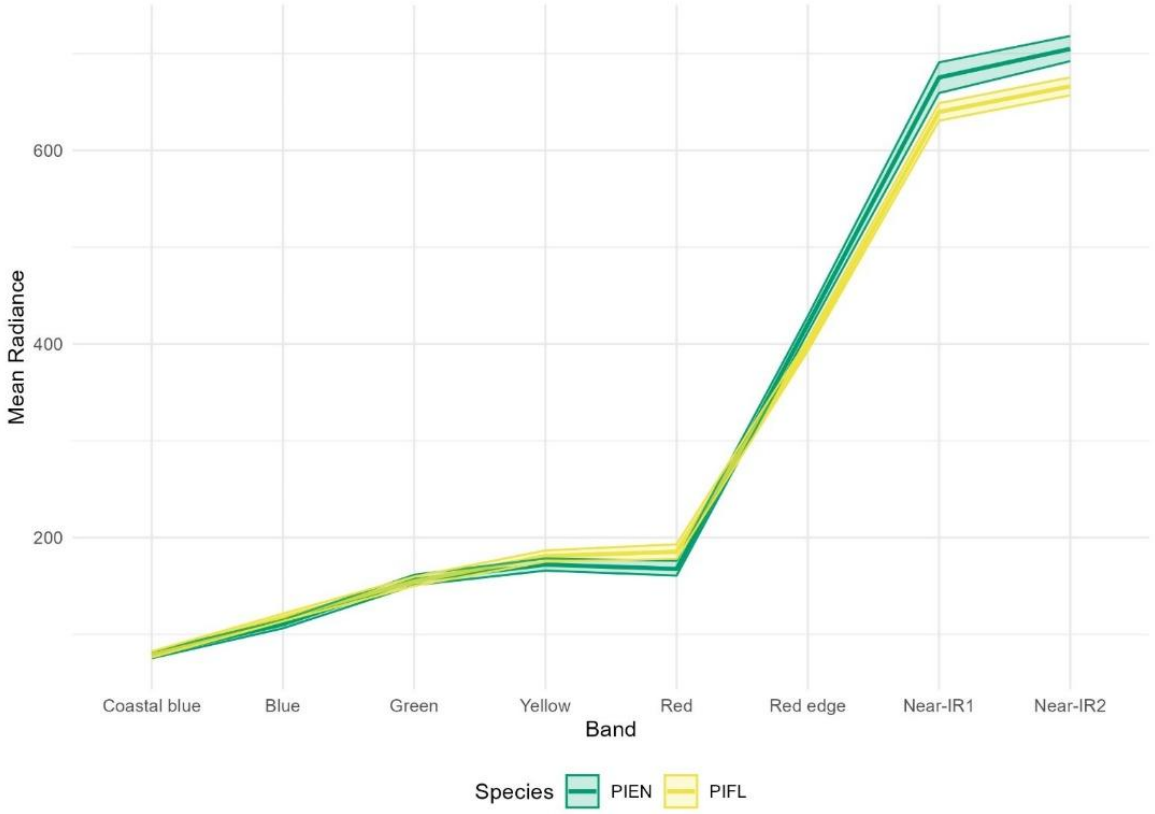

Figure 21. Surface radiance for PIEN vs. PIFL.

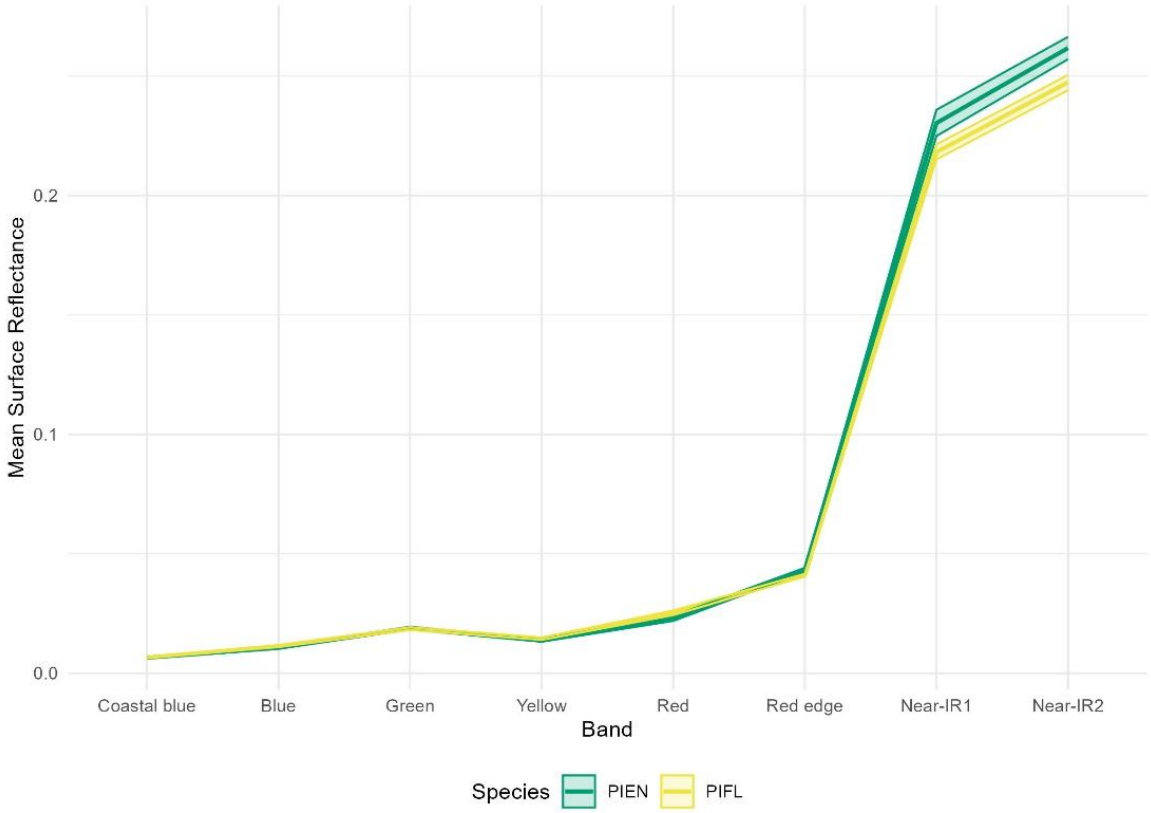

Figure 22. Surface reflectance for PIEN vs. PIFL.

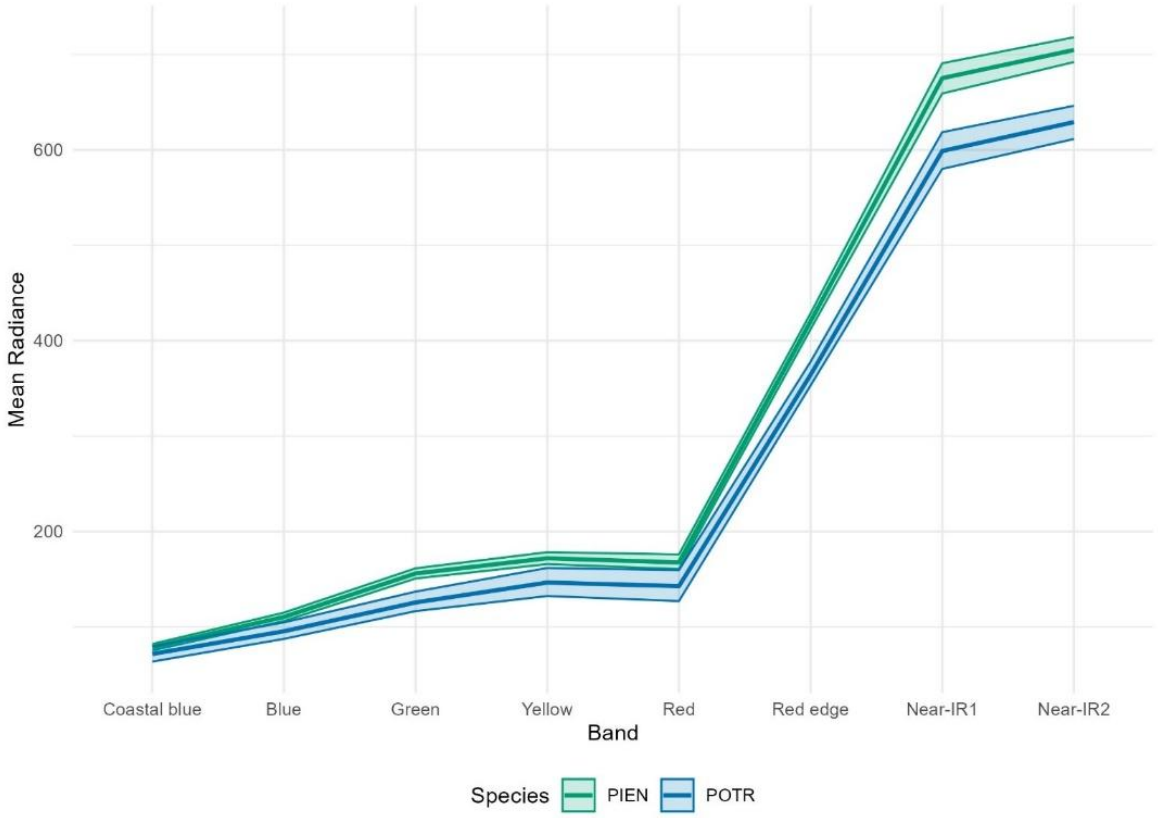

Figure 23. Surface radiance for PIEN vs. POTR.

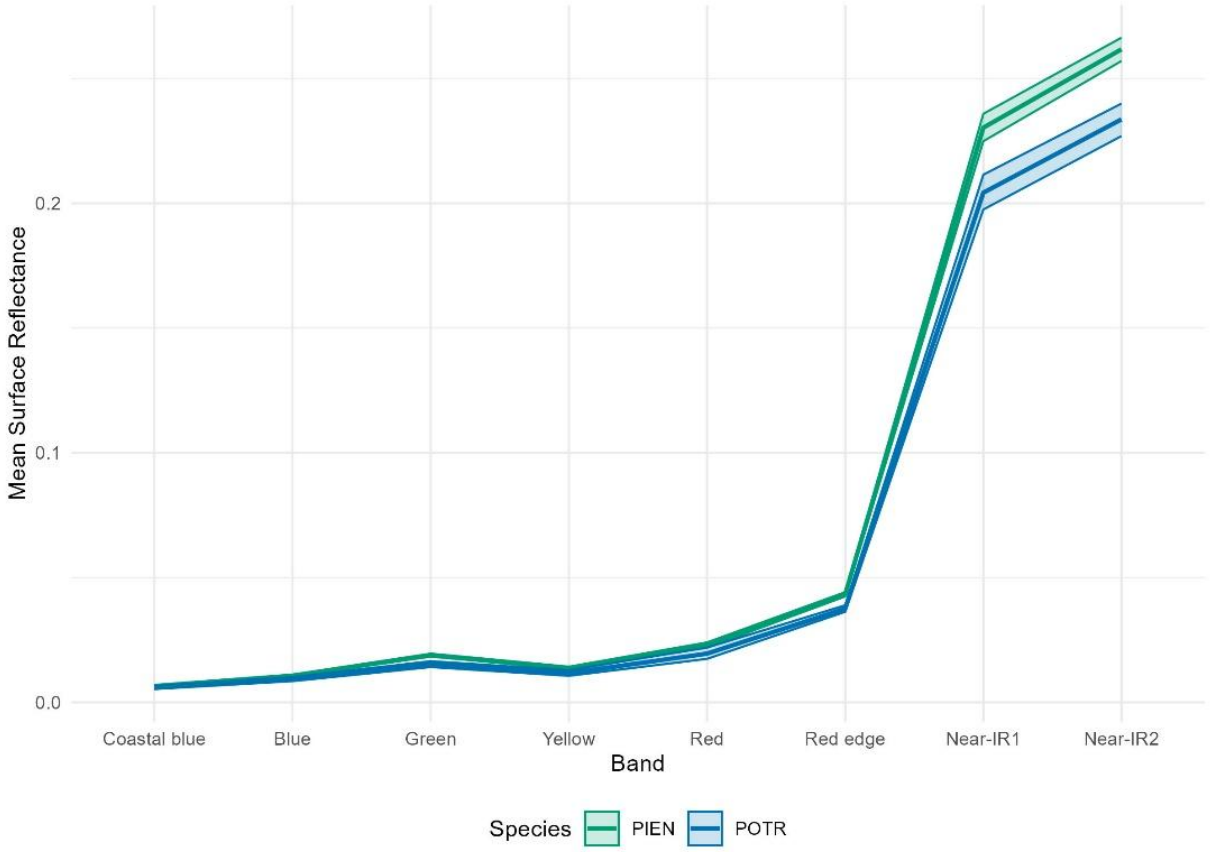

Figure 24. Surface reflectance for PIEN vs. POTR.

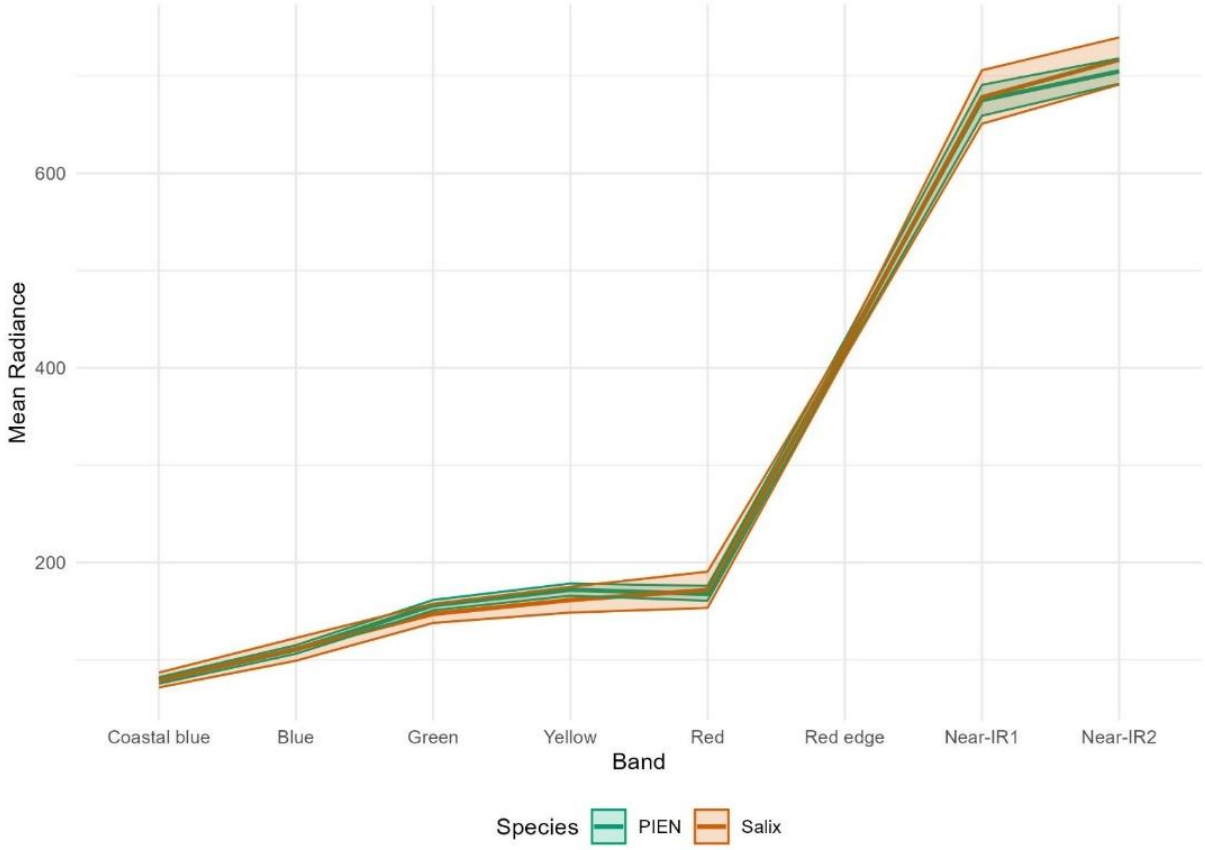

Figure 25. Surface radiance for PIEN vs. Salix.

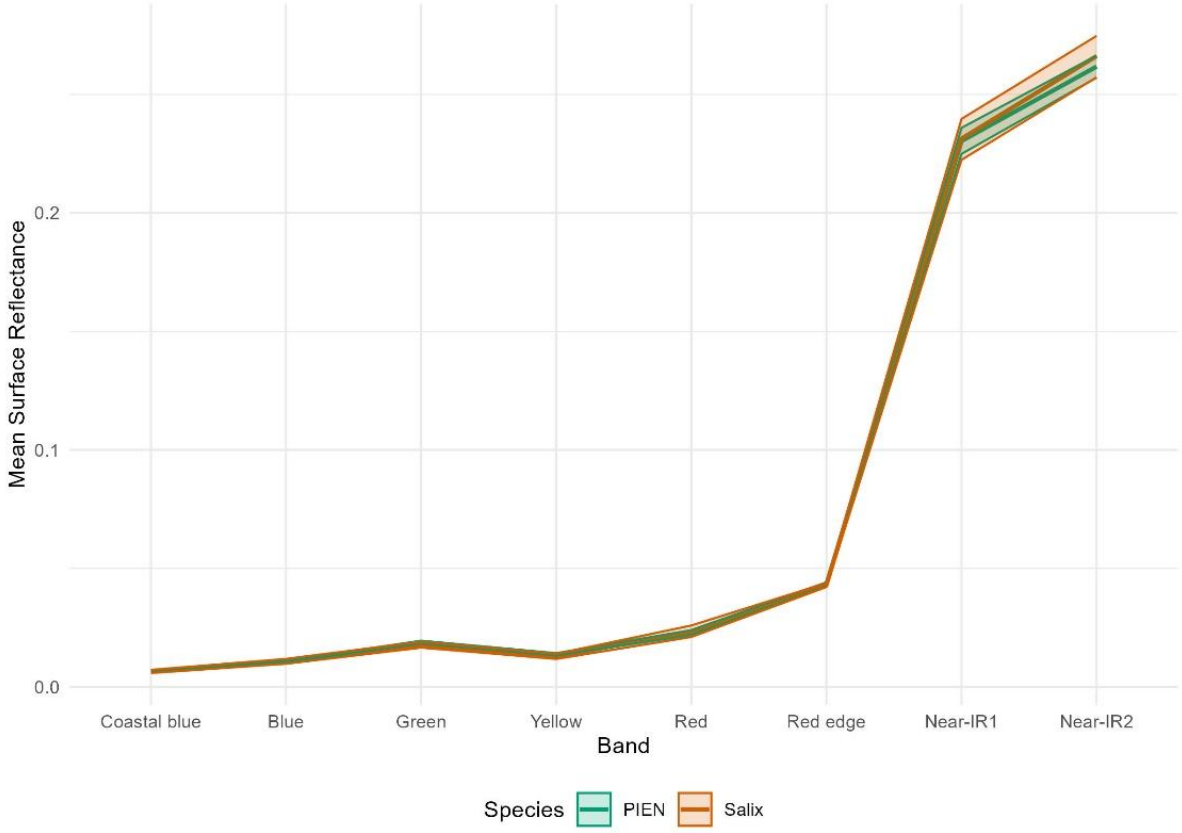

Figure 26. Surface reflectance for PIEN vs. Salix.

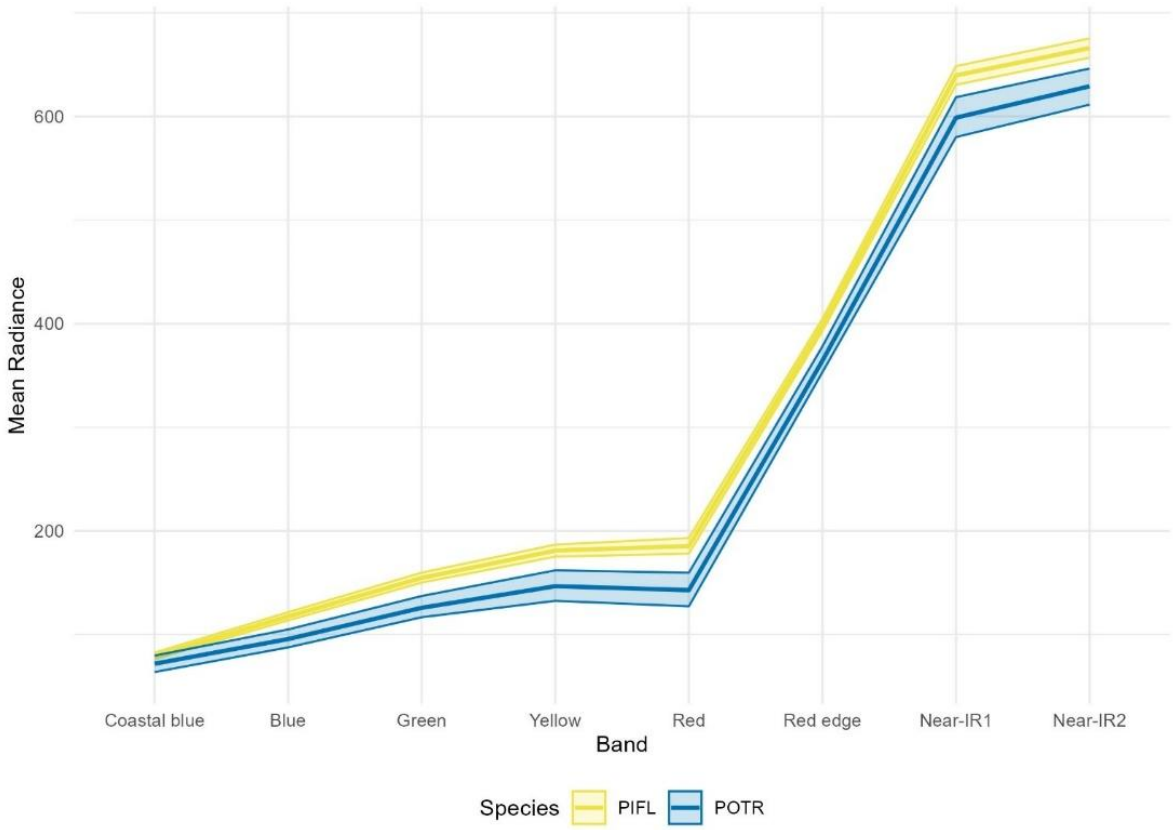

Figure 27. Surface radiance for PIFL vs. POTR.

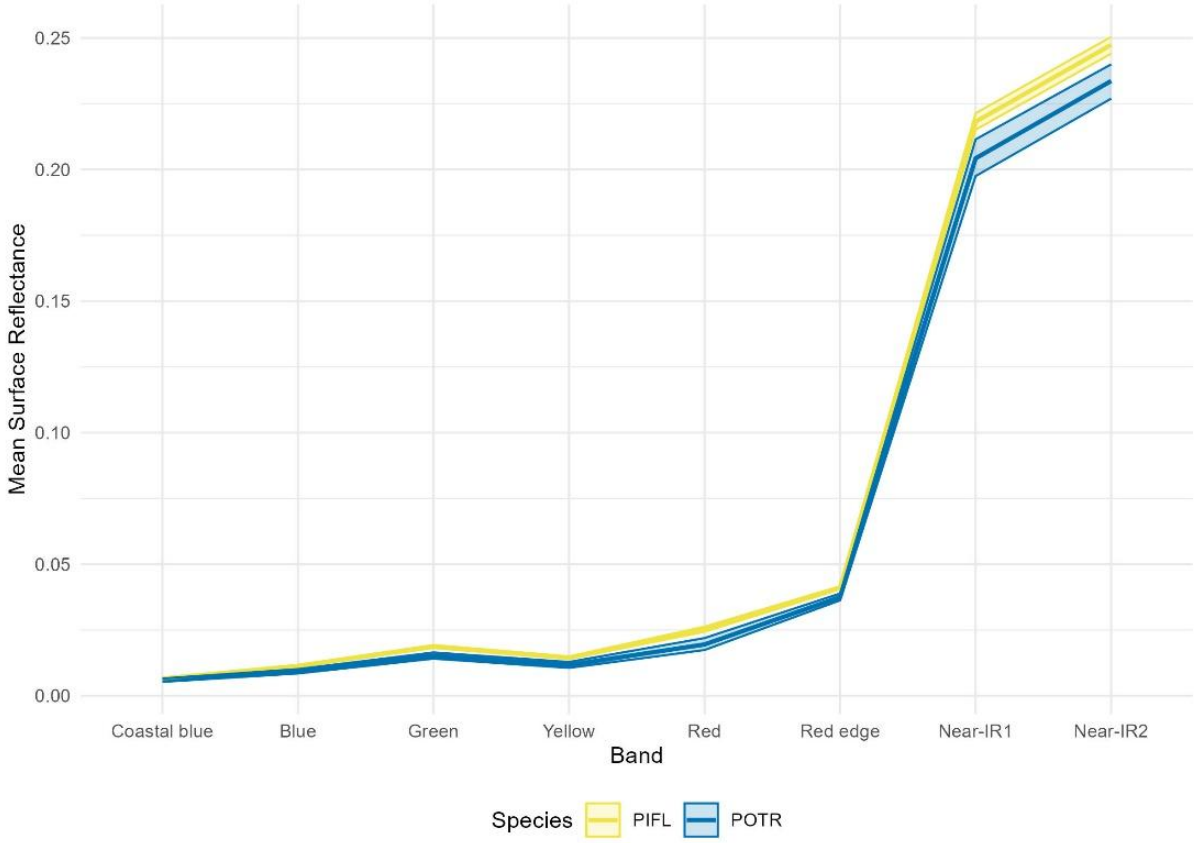

Figure 28. Surface reflectance for PIFL vs. POTR.

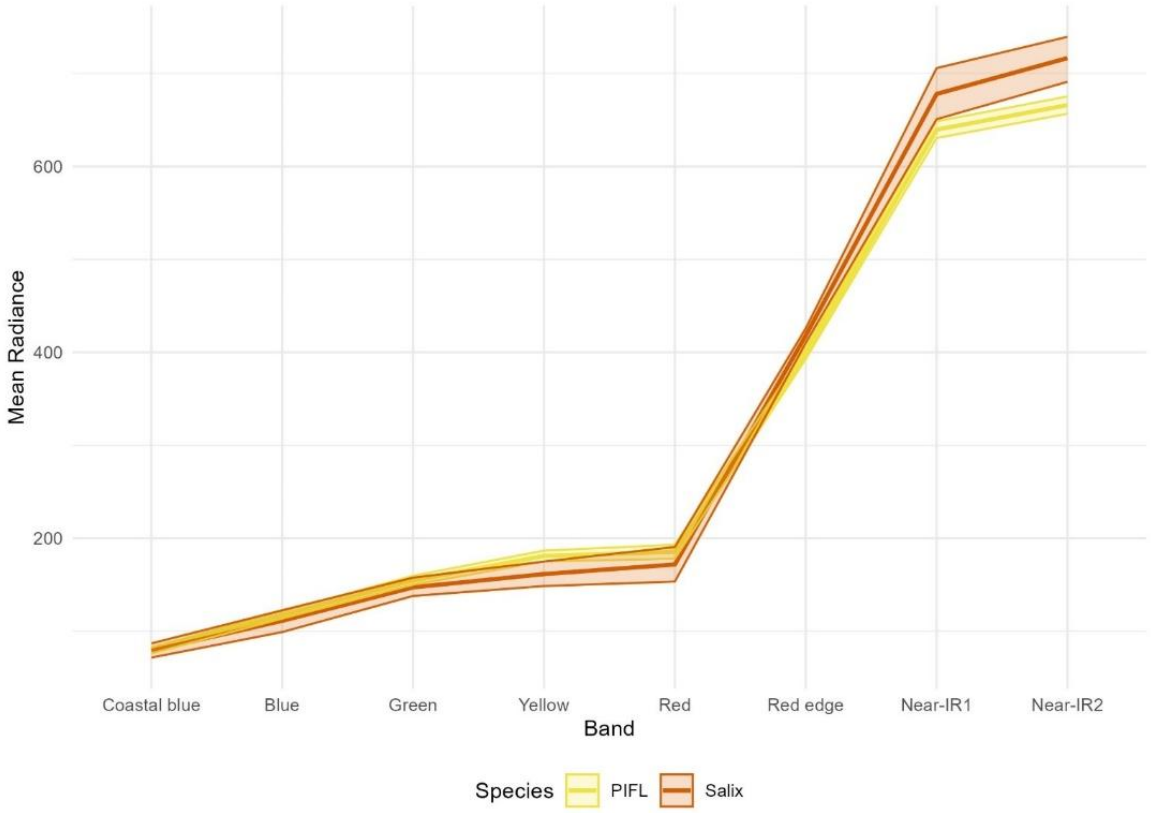

Figure 29. Surface radiance for PIFL vs. Salix.

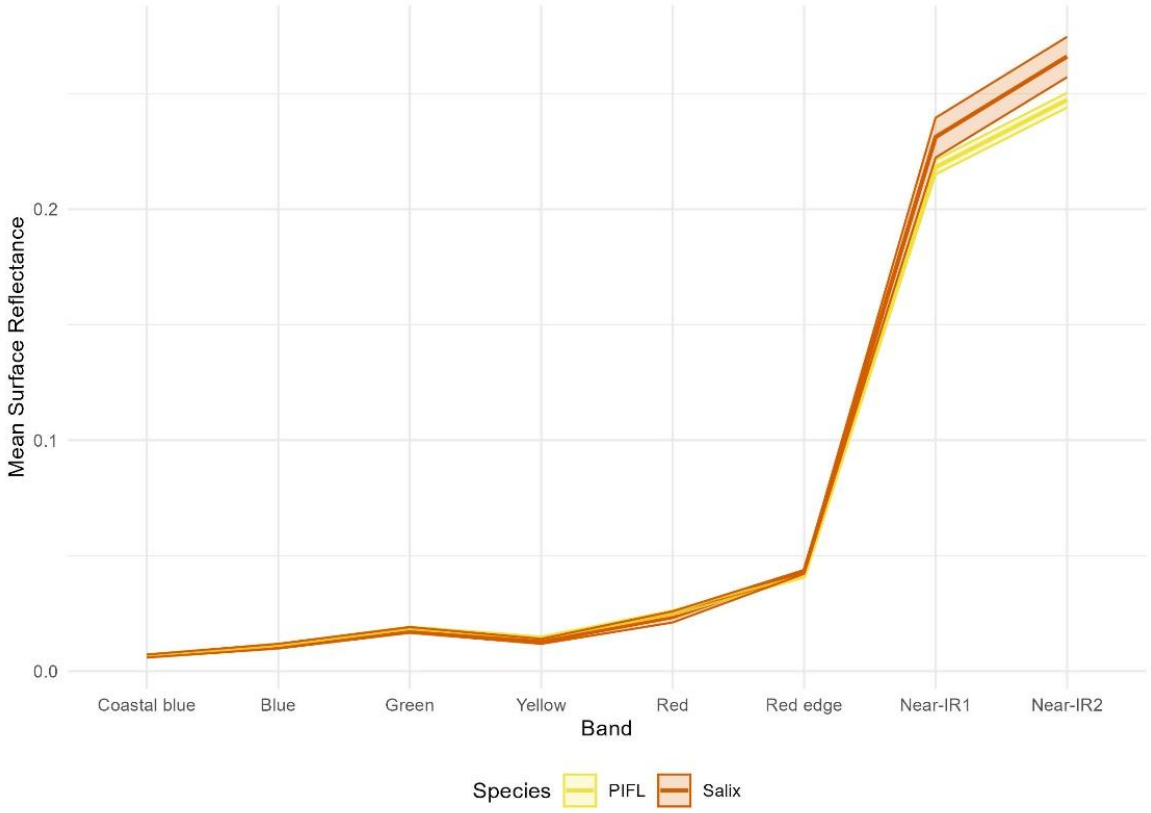

Figure 30. Surface reflectance for PIFL vs. Salix.

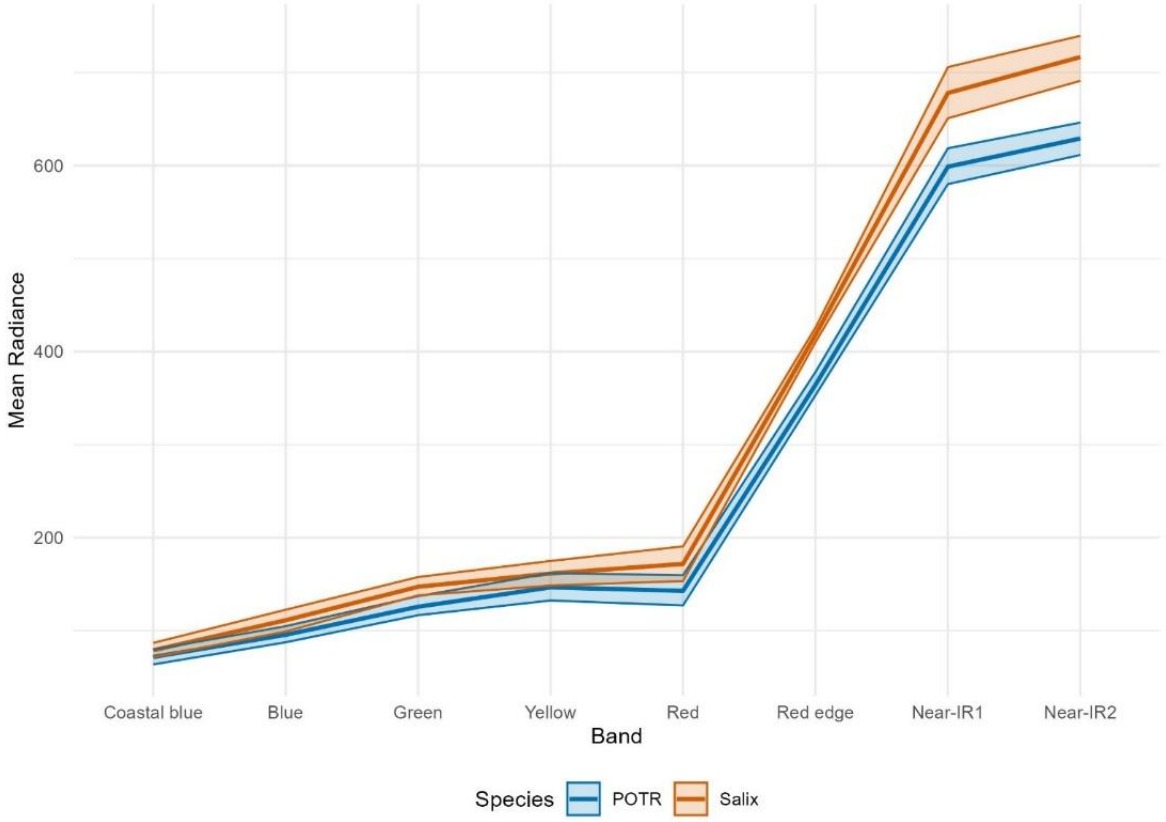

Figure 31. Surface radiance for POTR vs. Salix.

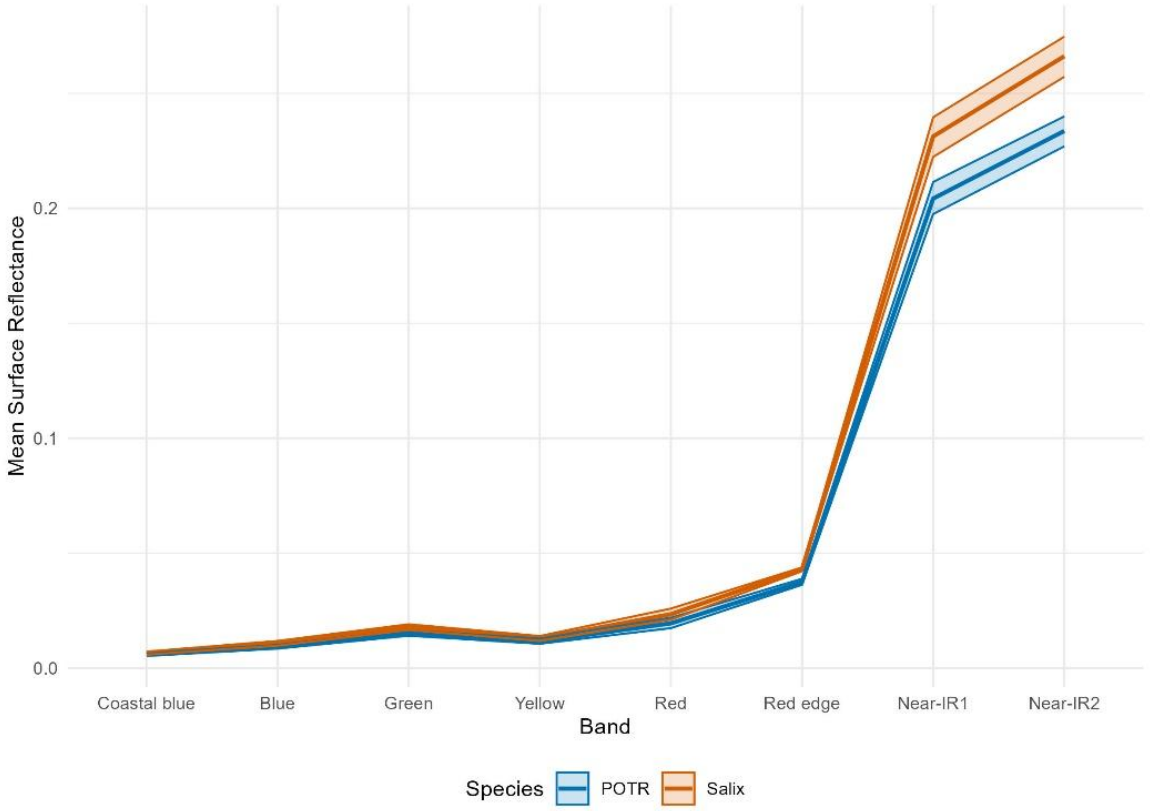

Figure 32. Surface reflectance for POTR vs. Salix.