Manuscript ID: ESSD-2025-613-RC

Title: GlobalGeoTree: A Multi-Granular Vision-Language Dataset for Global Tree Species Classification

Comments & Suggestions by REVIEWER#1 and our responses to them (shaded):

ESSD-2025-613-RC "*GlobalGeoTree: A Multi-Granular Vision-Language Dataset for Global Tree Species Classification*"

Comments by Reviewer#1:

General Comments: The authors present a global multi-modal dataset for tree species classification, integrating diverse data sources and offering both a large-scale pretraining dataset and a separate evaluation set. They also propose GeoTreeCLIP, a model that leverages hierarchical label structures and demonstrates improvements over baseline methods. The experimental setup is comprehensive, including comparisons with CLIP-style models and supervised learning approach. All code and data are publicly available.

Response from Authors: We thank the reviewer for the positive summary and for appreciating the scale and comprehensive nature of our work. We are grateful for the constructive comments, which have helped us significantly improve the manuscript's focus.

In this revision, we have addressed all specific points raised, with particular emphasis on strengthening the dataset validation and providing more granular performance analyses. Detailed responses to specific comments are provided below.

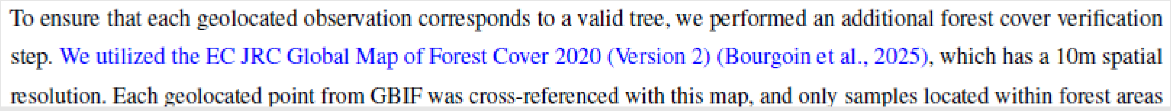
1. Dataset Construction

1.1 The authors use the JRC Forest Cover Map v1 for filtering. Given that version 2 has been publicly released with documented improvements, is there a reason for not using the updated version?

Response from Authors: We thank the reviewer for this careful observation and are pleased to have the opportunity to clarify our methodology. We confirm that we did utilize the JRC Global Map of Forest Cover 2020 Version 2 for our data filtering.

We apologize that the specific version number was not explicitly stated in the manuscript text, leading to this ambiguity. We selected Version 2 precisely because of its documented improvements in accuracy and spatial detail, ensuring the highest quality mask for our 2020 target year.

We have revised the text in **Section 3.1.3** to explicitly state "EC JRC Global Map of Forest Cover 2020 (Version 2)" to avoid any future confusion.



To ensure that each geolocated observation corresponds to a valid tree, we performed an additional forest cover verification step. We utilized the EC JRC Global Map of Forest Cover 2020 (Version 2) (Bourgoin et al., 2025), which has a 10m spatial resolution. Each geolocated point from GBIF was cross-referenced with this map, and only samples located within forest areas

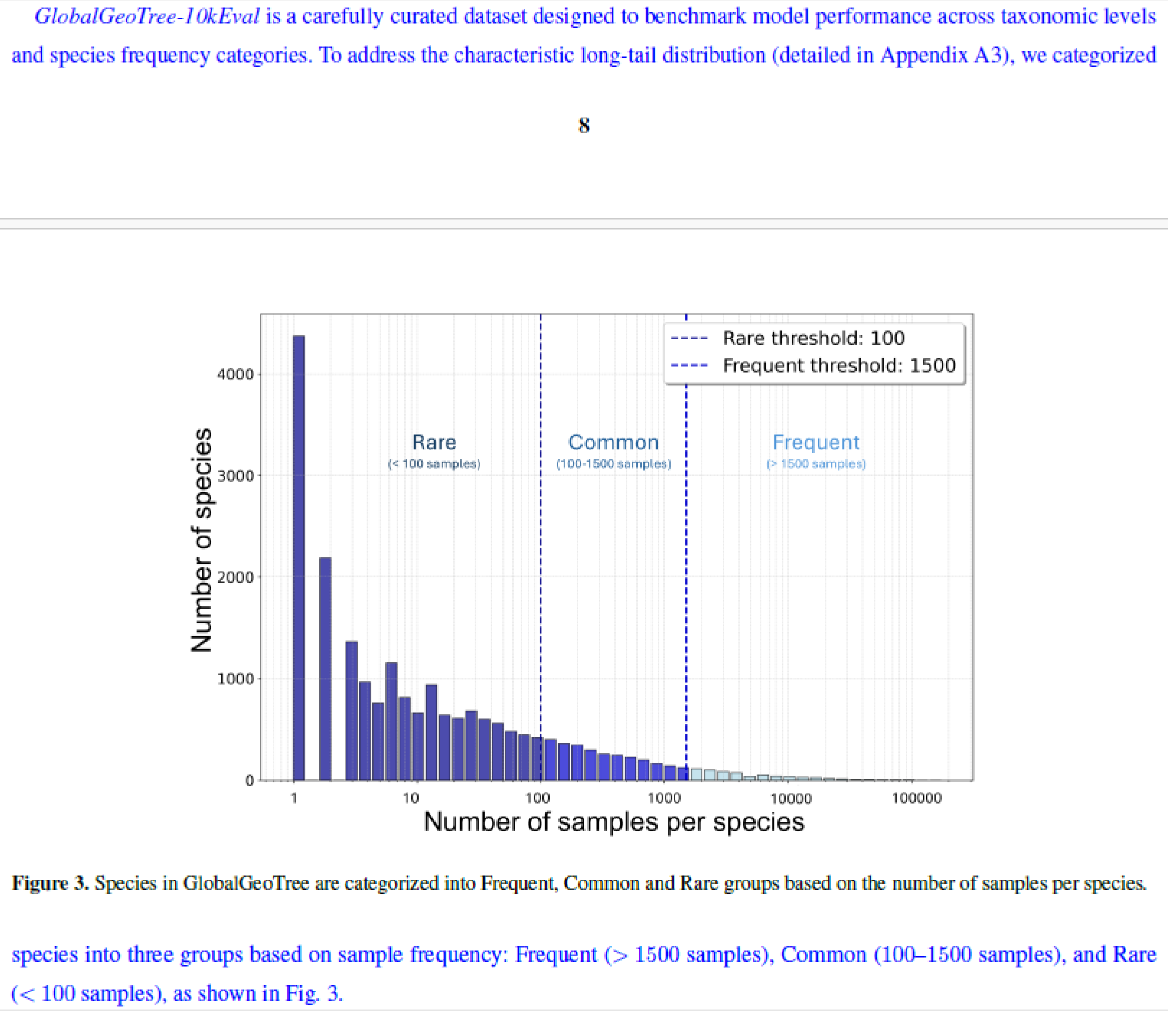
1.2 The GlobalGeoTree-10kEval set includes 90 species out of over 21,000. Could the authors clarify the selection criteria? Were any sampling or filtering strategies applied to ensure the reliability of the evaluation set, particularly given the inclusion of citizen science sources like iNaturalist?

Response from Authors: The selection of the 90 species followed a stratified random sampling strategy designed to represent the dataset's long-tail distribution. Species were categorized into Rare (< 100 samples), Common ($100 - 1500$ samples), and Frequent (> 1500 samples) groups based on sample frequency first. While the primary benchmark focuses on 90 species to allow for detailed analysis, we also constructed larger evaluation subsets containing 300 and 900 species to ensure evaluation robustness and diversity across broader taxonomic scales.

To ensure data reliability, we applied a strict multi-step filtering pipeline consistent with the general data curation process detailed in Section 3.1.2:

1. **Temporal Filtering:** Selecting only recent observations recorded between 2015 and 2024 to align with the improved accuracy of modern multi-constellation GNSS receivers.
2. **Observation Type:** Limiting data to human observation records to exclude fossil or specimen records, and exclusively utilize "Research-grade" records especially for iNaturalist, which guarantee high-quality metadata (date/location), include verifiable digital evidence (photo), and require taxonomic consensus from at least two independent contributors (often experts).
3. **Geospatial Quality:** Excluding records with geospatial issues as flagged by GBIF (e.g., country-coordinate mismatches).
4. **Presence Status:** Filtering for confirmed "present" occurrences.
5. **Precision Control:** Removing duplicate entries and observations with low geographic precision.

We have expanded **Section 3.4** to explicitly detail the stratification logic and the filtering steps.

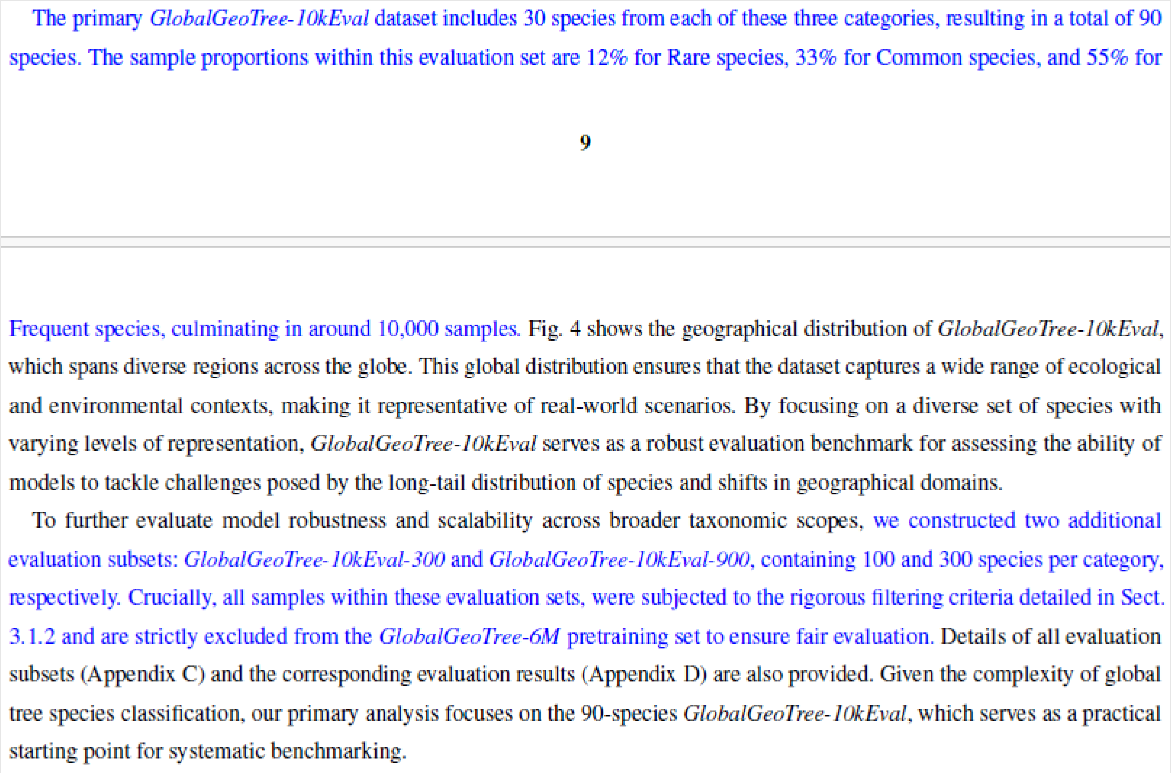


GlobalGeoTree-10kEval is a carefully curated dataset designed to benchmark model performance across taxonomic levels and species frequency categories. To address the characteristic long-tail distribution (detailed in Appendix A3), we categorized

8

Figure 3. Species in GlobalGeoTree are categorized into Frequent, Common and Rare groups based on the number of samples per species.

species into three groups based on sample frequency: Frequent (> 1500 samples), Common (100–1500 samples), and Rare (< 100 samples), as shown in Fig. 3.



The primary *GlobalGeoTree-10kEval* dataset includes 30 species from each of these three categories, resulting in a total of 90 species. The sample proportions within this evaluation set are 12% for Rare species, 33% for Common species, and 55% for

Frequent species, culminating in around 10,000 samples. Fig. 4 shows the geographical distribution of *GlobalGeoTree-10kEval*, which spans diverse regions across the globe. This global distribution ensures that the dataset captures a wide range of ecological and environmental contexts, making it representative of real-world scenarios. By focusing on a diverse set of species with varying levels of representation, *GlobalGeoTree-10kEval* serves as a robust evaluation benchmark for assessing the ability of models to tackle challenges posed by the long-tail distribution of species and shifts in geographical domains.

To further evaluate model robustness and scalability across broader taxonomic scopes, we constructed two additional evaluation subsets: *GlobalGeoTree-10kEval-300* and *GlobalGeoTree-10kEval-900*, containing 100 and 300 species per category, respectively. Crucially, all samples within these evaluation sets, were subjected to the rigorous filtering criteria detailed in Sect. 3.1.2 and are strictly excluded from the *GlobalGeoTree-6M* pretraining set to ensure fair evaluation. Details of all evaluation subsets (Appendix C) and the corresponding evaluation results (Appendix D) are also provided. Given the complexity of global tree species classification, our primary analysis focuses on the 90-species *GlobalGeoTree-10kEval*, which serves as a practical starting point for systematic benchmarking.

1.3 While the evaluation set is constructed as a separate test set, there appears to be no explicit validation process to assess its quality. Given the integration of heterogeneous data sources, some form of validation (manual or automated) would greatly enhance the trustworthiness and utility of the dataset.

Response from Authors: We fully agree that rigorous validation is crucial for an ESSD publication. While validating millions of global samples via fieldwork is infeasible, we have implemented a comprehensive **Technical Validation** framework to ensure data reliability.

1. Visual Validation Study: We conducted a visual inspection on a random sample of 300 locations from the *GlobalGeoTree-10kEval* dataset using Very High Resolution (VHR) imagery (e.g., Google Earth Maps). We found that **98.3%** of the samples clearly corresponded to forest or tree cover, while a small minority were ambiguous (e.g., forest edges). This high agreement confirms that our automated filtering pipeline effectively removes non-forest noise.

2. Transparent Community Validation (GEE App): To facilitate broad community inspection, we have deployed the entire GlobalGeoTree dataset on Google Earth Engine. This interactive tool allows users to zoom into any of the 6.3 million samples, overlay them on satellite imagery, and inspect detailed attributes. This transparency allows for continuous, crowdsourced qualitative validation (**Fig. 5**). The explorer is available at: `https://ee-yangm.projects.earthengine.app/view/globalgeotree-explorer`.

3. Cross-Validation with Independent Global Maps: We performed a stratified cross-validation against the *Copernicus Global Land Cover* (CGLS-LC100) product. Since CGLS provides discrete land cover information, we aligned our Level 0 labels (Functional Type) with CGLS classes at the leaf-type level (Broadleaf vs. Needleleaf). We sampled 10,000 points per continent (stratified) to mitigate regional bias. The results show strong agreement (average **80.38%** global consistency), particularly in South America and Africa (> 98%), demonstrating high taxonomic reliability at the leaf level.

4. Taxonomic Provenance: We rely on the provenance reliability of our multi-step rigorous filtering pipeline (detailed in Section 3) and verify that 100% of samples pass the GBIF taxonomic backbone check.

We have added a new **Section 4: Data Quality and Validation**, which details the VHR inspection, introduces the GEE Explorer App, and presents the cross-validation metrics.

4 Data Quality and Validation

To ensure the reliability of GlobalGeoTree, we implemented a multi-tiered validation framework combining fine-scale visual inspection, global-scale cross-verification with independent land cover products, and a public platform for community auditing.

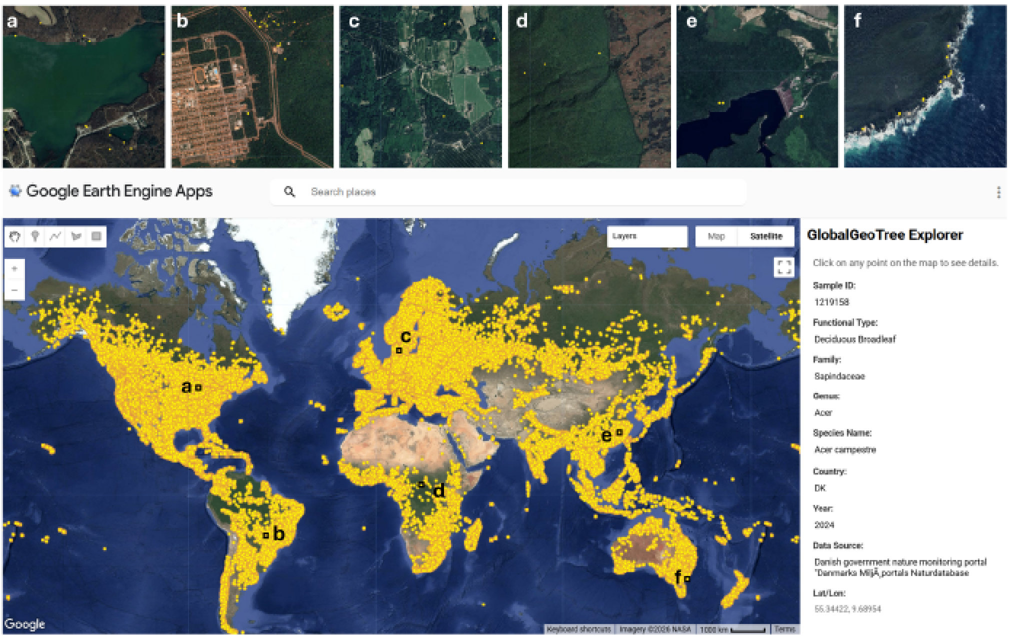
4.1 Visual Inspection and Community Validation

We first quantified land-cover accuracy through a visual validation study. A random subset of 300 locations was sampled from *GlobalGeoTree-10kEval*. For each sample, we retrieved Very High Resolution (VHR) satellite imagery to verify whether the geolocated point fell within forest or tree cover. The inspection confirmed that 98.3% of the samples were correctly located in forested areas, while a small minority were ambiguous (e.g., forest edges), validating the efficacy of our automated filtering pipeline.

To further enhance transparency and enable large-scale qualitative validation by the research community, we developed the GlobalGeoTree Explorer, a web-based application hosted on Google Earth Engine (Fig. 5). This tool allows users to visualize the spatial distribution of all 6.3 million samples, overlay them on high-resolution satellite basemaps, and inspect detailed taxonomic attributes for any individual point. The application is publicly accessible at https://ee-yangm.projects.earthengine.app/view/globalgeotree-explorer.

4.2 Cross-Validation with Global Land Cover Products

To validate taxonomic consistency at the functional type level, we performed a cross-comparison against the Copernicus Global Land Cover Layers (CGLS-LC100, Collection 3) (Buchhorn et al., 2020). Since no global dataset currently provides



GlobalGeoTree Explorer

Click on any point on the map to see details.

Sample ID:
1219158

Functional Type:
Deciduous Broadleaf

Family:
Sapindaceae

Genus:
Acer

Species Name:
Acer campestre

Country:
DK

Year:
2024

Data Source:
Danish government nature monitoring portal
'Danmarks MiljÃ¸portals Naturdatabase

Lat/Lon:
55.34422, 9.68954

Figure 5. The GlobalGeoTree Explorer App. The interface allows users to visualize the global distribution of tree samples (bottom map) and inspect individual points against high-resolution satellite imagery (top panels a-f). Detailed attributes for selected samples are displayed in the sidebar, facilitating transparent community validation.

high-resolution distribution maps at the species or genus level, the CGLS-LC100 product represents the most detailed global baseline available, offering discrete land cover layers at 100m resolution that distinguish between broadleaf and needleleaf forest types.

We aligned the `level0` labels of GlobalGeoTree with CGLS classes by aggregating them into two primary leaf types: Needleleaf and Broadleaf. To ensure global representativeness and mitigate regional data density biases, we employed a stratified sampling strategy, selecting 10,000 samples per continent.

Figure 6 presents the results of this cross-validation. The confusion matrix (left) reveals an overall agreement of 80.38% at the leaf level. We observed high consistency for Broadleaf samples (81%), while some Needleleaf samples (22%) showed discrepancies with the coarser 100m product, likely due to mixed pixels in transition zones. The regional breakdown (right) demonstrates exceptionally high agreement in South America (99.3%) and Africa (98.7%), where broadleaf forests dominate. Lower agreement rates in North America and Europe are expected given the higher prevalence of mixed forests and the resolution gap between our point-based data and the 100m validation raster.

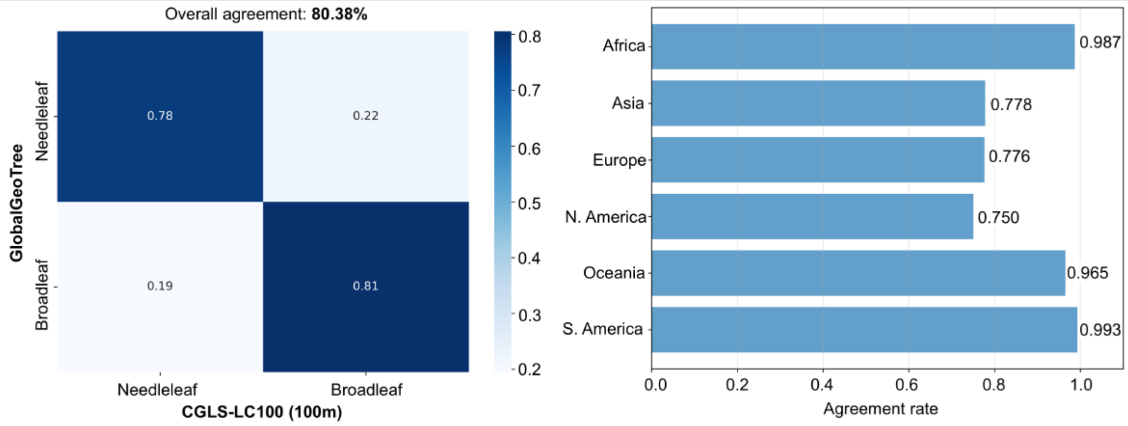


Figure 6. Cross-validation results against the CGLS-LC100 product. **Left:** Normalized confusion matrix at the Leaf Level (Needleleaf vs. Broadleaf), showing an overall agreement of 80.38%. **Right:** Agreement rates broken down by continent, highlighting strong consistency in tropical regions.

4.3 Taxonomic Reliability

Beyond spatial validation, taxonomic reliability is ensured through our strict data provenance pipeline. We exclusively utilize observations which require community consensus for species identification, and 100% of these samples have passed the GBIF taxonomic backbone check (GBIF Secretariat, 2023).

2. Model (GeoTreeCLIP)

2.1 The authors attribute the performance improvements of GeoTreeCLIP to domain-specific pretraining. However, it's difficult to isolate the effects of pretraining alone, as other baseline models lack temporal fusion and may differ in how auxiliary data are handled. A more controlled ablation or discussion would strengthen this claim.

Response from Authors: We agree that disentangling the contributions of the model architecture from the pretraining paradigm is essential for a rigorous analysis. To strictly isolate these effects, we rely on three controlled comparisons:

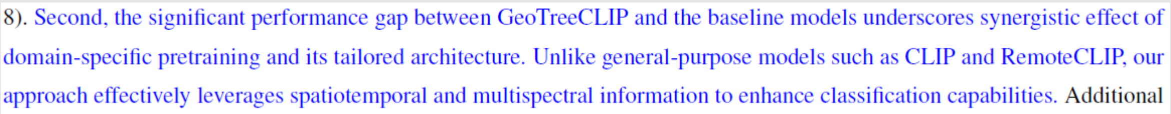
1. **Architecture Control (SupervisedGeoTree vs. GeoTreeCLIP):** We compared GeoTreeCLIP against a SupervisedGeoTree baseline (Section 6.1). Both models use the exact same architecture (Visual Encoder + MLP for auxiliary data). The only difference is the learning objective (Contrastive vs. Supervised Cross-Entropy). The superior performance of GeoTreeCLIP (Zero-shot) and its efficiency in Few-shot settings strongly suggest that the *VLM pretraining paradigm* on our dataset effectively learns semantic relationships that supervised learning misses.

2. **Pretraining Data Control (General RS VLM vs. GeoTreeCLIP):** In Section 5.3 and Appendix E, we compare our model against RemoteCLIP, SkyCLIP-50 and CLIP-laion-RS. These models employ a similar vision-language framework but are pretrained on generic remote sensing image-text pairs. Their significantly lower accuracy on the tree species classification task highlights the importance of the domain-specific data (*GlobalGeoTree-6M*), rather than the VLM approach itself.

3. **Input Modality Ablation:** As suggested, we present an ablation study in Section 6.3 (Table 7) that systematically removes auxiliary variables and multi-spectral bands. These results quantify the specific gains attributable to the multimodal architecture, confirming that while pre-

> training is foundational, the integration of spatiotemporal and environmental context is equally necessary.

We appreciate the reviewer pointing out this potential ambiguity. We agree that the performance gap is driven by a combination of two factors: the **specialized architecture** (which can ingest time-series and auxiliary data) and the **domain-specific pretraining data**. It was not our intention to attribute the improvement solely to pretraining. We have revised the discussion in **Section 5.3.1** to explicitly synthesize these findings.



8). Second, the significant performance gap between GeoTreeCLIP and the baseline models underscores synergistic effect of domain-specific pretraining and its tailored architecture. Unlike general-purpose models such as CLIP and RemoteCLIP, our approach effectively leverages spatiotemporal and multispectral information to enhance classification capabilities. Additional

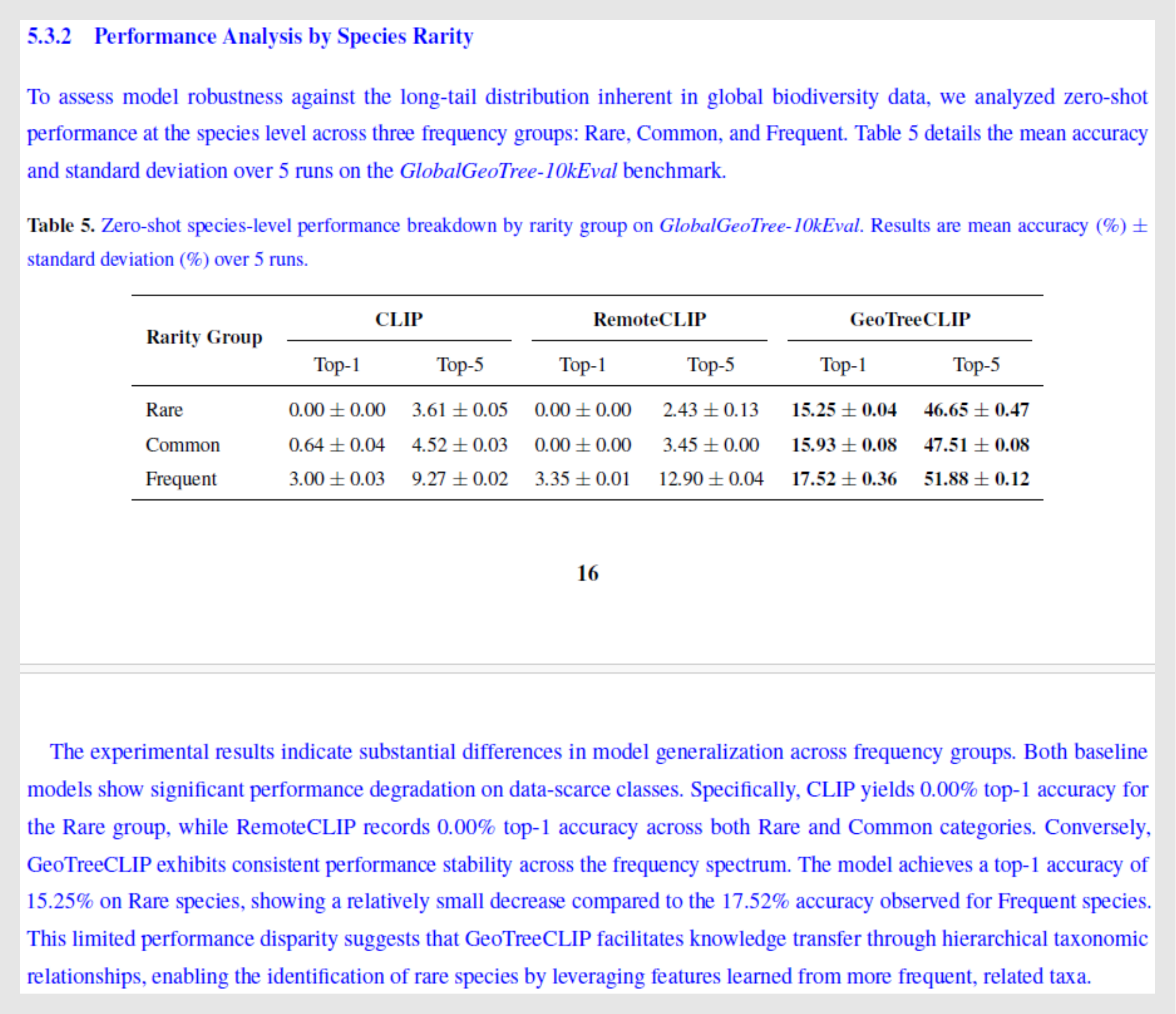
3. Evaluation Metrics and Reporting

3.1 The paper mentions addressing class imbalance by grouping species into frequent, common, and rare categories. However, results are not reported per group. Including group-specific performance would align with common practices in imbalanced classification tasks.

Response from Authors: We fully agree that analyzing performance across frequency groups is essential for evaluating model robustness on long-tail distributions. We have performed a zero-shot evaluation on the *GlobalGeoTree-10kEval* dataset, breaking down the species-level accuracy into Rare, Common, and Frequent categories.

The results, now detailed in **Table 5**, show distinct performance patterns. While baseline models such as CLIP and RemoteCLIP yield 0.00% top-1 accuracy on the Rare group, GeoTreeCLIP achieves **15.25%** top-1 accuracy. Furthermore, the performance drop from Frequent (17.52%) to Rare (15.25%) species in GeoTreeCLIP is relatively small compared to the baselines. This suggests that the model's pretraining on hierarchical taxonomic labels enables effective feature transfer from frequent to rare classes.

We have added a new subsection **"Performance Analysis by Species Rarity"** in **Section 5.3**, which includes a table presenting the mean top-1 and top-5 accuracy for each rarity group across all three models.



5.3.2 Performance Analysis by Species Rarity

To assess model robustness against the long-tail distribution inherent in global biodiversity data, we analyzed zero-shot performance at the species level across three frequency groups: Rare, Common, and Frequent. Table 5 details the mean accuracy and standard deviation over 5 runs on the *GlobalGeoTree-10kEval* benchmark.

Table 5. Zero-shot species-level performance breakdown by rarity group on *GlobalGeoTree-10kEval*. Results are mean accuracy (%) ± standard deviation (%) over 5 runs.

Rarity Group	CLIP		RemoteCLIP		GeoTreeCLIP	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Rare	0.00 ± 0.00	3.61 ± 0.05	0.00 ± 0.00	2.43 ± 0.13	**15.25 ± 0.04**	**46.65 ± 0.47**
Common	0.64 ± 0.04	4.52 ± 0.03	0.00 ± 0.00	3.45 ± 0.00	**15.93 ± 0.08**	**47.51 ± 0.08**
Frequent	3.00 ± 0.03	9.27 ± 0.02	3.35 ± 0.01	12.90 ± 0.04	**17.52 ± 0.36**	**51.88 ± 0.12**

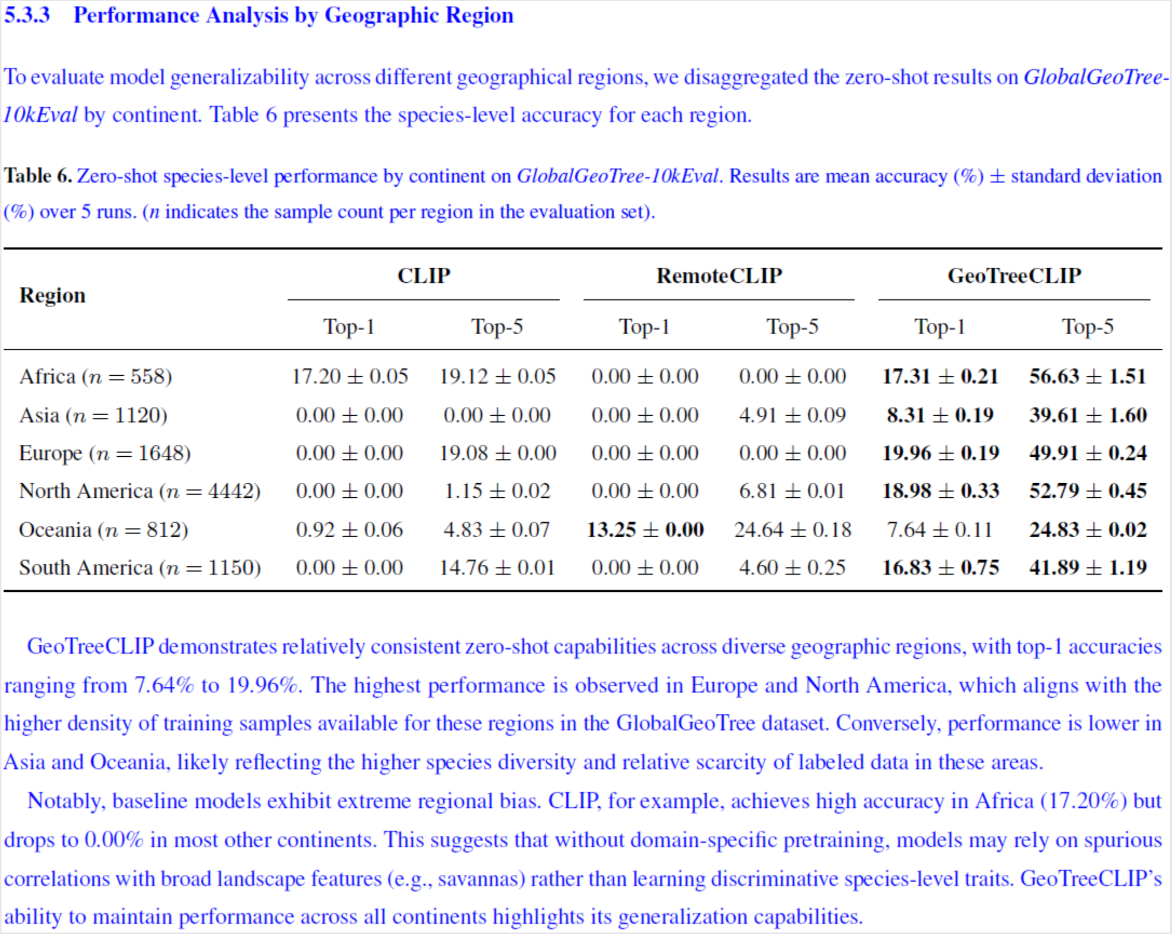
The experimental results indicate substantial differences in model generalization across frequency groups. Both baseline models show significant performance degradation on data-scarce classes. Specifically, CLIP yields 0.00% top-1 accuracy for the Rare group, while RemoteCLIP records 0.00% top-1 accuracy across both Rare and Common categories. Conversely, GeoTreeCLIP exhibits consistent performance stability across the frequency spectrum. The model achieves a top-1 accuracy of 15.25% on Rare species, showing a relatively small decrease compared to the 17.52% accuracy observed for Frequent species. This limited performance disparity suggests that GeoTreeCLIP facilitates knowledge transfer through hierarchical taxonomic relationships, enabling the identification of rare species by leveraging features learned from more frequent, related taxa.

3.2 Given the global scope of the dataset and the known regional biases, regional performance breakdowns would be informative and important for understanding model generalizability.

Response from Authors: This is a valuable addition to assess generalizability given the uneven geographic distribution of biodiversity data. We have conducted a continent-wise zero-shot performance analysis on the *GlobalGeoTree-10kEval* dataset.

The results (detailed in Table 6) indicate that GeoTreeCLIP maintains functional zero-shot capabilities across all continents, achieving its highest top-1 accuracy in Europe (19.96%) and North America (18.98%), which correspond to regions with higher training data density. Performance is lower in regions with higher species diversity or lower sample counts, such as Asia (8.31%) and Oceania (7.64%). In contrast, baseline models show extreme regional instability; for instance, CLIP achieves 17.20% accuracy in Africa but 0.00% in most other regions, suggesting it may be overfitting to specific landscape features rather than learning generalized species representations.

We have added a new subsection **"Performance Analysis by Geographic Region"** in **Section 5.3**, accompanied by a detailed table breaking down the top-1 and top-5 accuracies for each continent.



5.3.3 Performance Analysis by Geographic Region

To evaluate model generalizability across different geographical regions, we disaggregated the zero-shot results on *GlobalGeoTree-10kEval* by continent. Table 6 presents the species-level accuracy for each region.

Table 6. Zero-shot species-level performance by continent on *GlobalGeoTree-10kEval*. Results are mean accuracy (%) ± standard deviation (%) over 5 runs. (n indicates the sample count per region in the evaluation set).

Region	CLIP		RemoteCLIP		GeoTreeCLIP	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Africa ($n = 558$)	17.20 ± 0.05	19.12 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	**17.31 ± 0.21**	**56.63 ± 1.51**
Asia ($n = 1120$)	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	4.91 ± 0.09	**8.31 ± 0.19**	**39.61 ± 1.60**
Europe ($n = 1648$)	0.00 ± 0.00	19.08 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**19.96 ± 0.19**	**49.91 ± 0.24**
North America ($n = 4442$)	0.00 ± 0.00	1.15 ± 0.02	0.00 ± 0.00	6.81 ± 0.01	**18.98 ± 0.33**	**52.79 ± 0.45**
Oceania ($n = 812$)	0.92 ± 0.06	4.83 ± 0.07	**13.25 ± 0.00**	24.64 ± 0.18	7.64 ± 0.11	**24.83 ± 0.02**
South America ($n = 1150$)	0.00 ± 0.00	14.76 ± 0.01	0.00 ± 0.00	4.60 ± 0.25	**16.83 ± 0.75**	**41.89 ± 1.19**

GeoTreeCLIP demonstrates relatively consistent zero-shot capabilities across diverse geographic regions, with top-1 accuracies ranging from 7.64% to 19.96%. The highest performance is observed in Europe and North America, which aligns with the higher density of training samples available for these regions in the GlobalGeoTree dataset. Conversely, performance is lower in Asia and Oceania, likely reflecting the higher species diversity and relative scarcity of labeled data in these areas.

Notably, baseline models exhibit extreme regional bias. CLIP, for example, achieves high accuracy in Africa (17.20%) but drops to 0.00% in most other continents. This suggests that without domain-specific pretraining, models may rely on spurious correlations with broad landscape features (e.g., savannas) rather than learning discriminative species-level traits. GeoTreeCLIP's ability to maintain performance across all continents highlights its generalization capabilities.

Additional comments: The authors' effort in assembling such a large-scale, publicly available dataset and developing a strong benchmark model is highly appreciated. However, since this is a data description paper, the dataset itself should be the focal point. At present, the lack of validation for datasets is a significant limitation. While the work offers valuable contributions for machine learning research, particularly within benchmark or workshop tracks at venues like CVPR or NeurIPS, it may not yet meet the expectations for a journal like ESSD, which prioritizes data quality.

Response from Authors: We sincerely thank the reviewer for this candid and critical assessment. We fully agree that for Earth System Science Data, the primary contribution must be the dataset's quality, reliability, and transparency, rather than the machine learning model itself.

In response to this valid concern, we have fundamentally restructured and expanded the manuscript to place the dataset at the center. Specifically, we have implemented a rigorous **Data Quality and Validation** framework (new Section 4) that goes beyond standard ML metrics:

1. **Visual Validation:** We performed a stratified visual inspection of 300 samples using VHR imagery, confirming a 98.3% land-cover accuracy.

2. **Global Cross-Validation:** We cross-referenced our 6.3 million samples against the independent global CGLS-LC100 product, demonstrating 80.38% overall agreement at the leaf-type level.

3. **Transparency:** We developed and released a **GlobalGeoTree Explorer App** on Google Earth Engine, allowing the scientific community to audit the quality of all individual samples directly.

4. **Granular Analysis:** We added detailed breakdowns of dataset performance by Species Rarity

(Table 5) and Geographic Region (Table 6) to transparently reveal the dataset's strengths and limitations.

With these additions, we position GlobalGeoTree not just as a playground for ML models, but as a verified, quality-controlled product for the Earth Science community.

Comments & Suggestions by REVIEWER#2
and our responses to them (shaded):

ESSD-2025-613-RC "*GlobalGeoTree: A Multi-Granular Vision-Language Dataset for Global Tree Species Classification*"

Comments by Reviewer#2:

General Comments: This manuscript presents a large-scale, multimodal dataset for global tree species mapping and provides baseline models built on vision–language architectures. The study is novel and well-executed, with clear potential impact. I particularly appreciate the innovative data integration and modeling approach. I think this paper has a strong contribution to the AI4forest and AI4ecology communities. I have a few points for discussion regarding some methodological choices, data integration strategies, validation procedures, and practical applicability, which could benefit from further clarification.

Response from Authors: We deeply appreciate the reviewer's positive assessment and the recognition of our work's contribution to the AI4forest and AI4ecology communities. The insightful comments regarding methodological choices, validation independence, and practical deployment are highly valuable. We have addressed each point below, adding clarifications to the manuscript and expanding the discussion on real-world applicability.

Specific Comments

1. MLP Branch vs. Extended Raster Cube

The authors convert all non–Sentinel-2 environmental layers (e.g., bioclimatic, soil, and SRTM data) into single scalar values and feed them into an additional MLP branch. Given that these auxiliary layers are nearly constant across each Sentinel-2 patch, it remains unclear whether a dedicated MLP module is efficient. A more straightforward and parameter-efficient alternative would be to treat these raster layers as additional bands and integrate them directly into the Sentinel-2 data cube. I suggest the authors simply discuss: Why not merge these coarse-resolution raster layers into the Sentinel-2 cube and encode them jointly with the visual encoder? What are the advantages and disadvantages of the current "scalar + MLP" design compared with the "extended raster cube" approach? How do these two strategies differ in terms of model capacity, parameter efficiency, and representational effectiveness?

Response from Authors: This is a critical architectural decision. We chose the **"Scalar + MLP"** design over the "Extended Raster Cube" approach for three key reasons, which we have now clarified in **Section 4.1**:

1. **Resolution Mismatch & Redundancy:** Sentinel-2 data is at 10m resolution, while our auxiliary variables are much coarser (e.g., WorldClim at ∼1km, SoilGrids at 250m). Merging them would require upsampling the environmental data by factor of 100x or 25x. This introduces massive spatial redundancy—a 5x5 Sentinel-2 patch would effectively see a "flat" constant value for these environmental channels. Using a CNN/ViT to process this constant value is computationally wasteful.

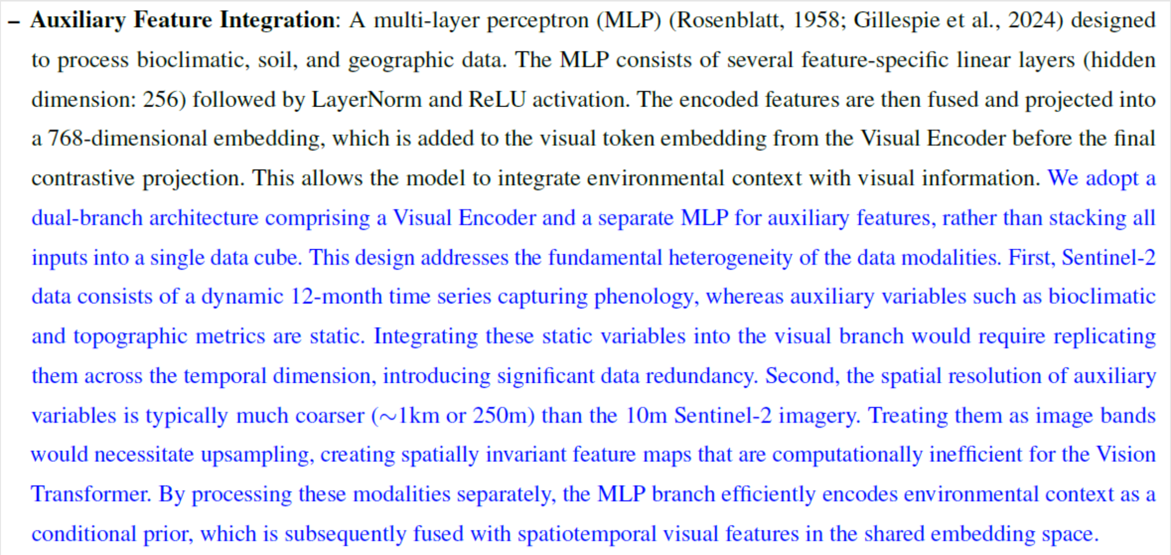
2. **Temporal Mismatch (Static vs. Dynamic):** Sentinel-2 data is a 12-month time series designed to capture phenological changes, whereas the auxiliary variables (e.g., elevation, annual rainfall) are static. Merging them into the data cube would require replicating the static variables 12 times across the temporal dimension, introducing significant redundancy and memory

overhead without adding information.

3. **Semantic Role:** Physically, these variables (temperature, elevation, soil pH) act as **Contextual Priors** (conditioning the probability of species occurrence) rather than **Visual Features** (like leaf texture or canopy shape). An MLP is the natural architecture to encode such structured tabular data, whereas Visual Encoders are optimized for spatial patterns.

4. **Flexibility:** The MLP branch allows the model to work even if imagery is missing or quality is poor (e.g., inference using only environmental priors), or to easily swap/add new environmental variables without retraining the entire visual backbone.

We have added a discussion in **Section 4.1** justifying the "Dual-Branch" architecture based on computational efficiency and the semantic distinction between visual texture and environmental context.



– **Auxiliary Feature Integration**: A multi-layer perceptron (MLP) (Rosenblatt, 1958; Gillespie et al., 2024) designed to process bioclimatic, soil, and geographic data. The MLP consists of several feature-specific linear layers (hidden dimension: 256) followed by LayerNorm and ReLU activation. The encoded features are then fused and projected into a 768-dimensional embedding, which is added to the visual token embedding from the Visual Encoder before the final contrastive projection. This allows the model to integrate environmental context with visual information. We adopt a dual-branch architecture comprising a Visual Encoder and a separate MLP for auxiliary features, rather than stacking all inputs into a single data cube. This design addresses the fundamental heterogeneity of the data modalities. First, Sentinel-2 data consists of a dynamic 12-month time series capturing phenology, whereas auxiliary variables such as bioclimatic and topographic metrics are static. Integrating these static variables into the visual branch would require replicating them across the temporal dimension, introducing significant data redundancy. Second, the spatial resolution of auxiliary variables is typically much coarser (∼1km or 250m) than the 10m Sentinel-2 imagery. Treating them as image bands would necessitate upsampling, creating spatially invariant feature maps that are computationally inefficient for the Vision Transformer. By processing these modalities separately, the MLP branch efficiently encodes environmental context as a conditional prior, which is subsequently fused with spatiotemporal visual features in the shared embedding space.

2. Validation Independence

The validation set is derived from the same volunteer-contributed datasets (e.g., GBIF, iNaturalist) that were used for training. These datasets share similar sources of error and observation bias. I recommend incorporating independent validation sources, such as national forest inventories, ecological monitoring networks, or field-based regional datasets. If additional data cannot be included, a brief discussion is needed on how shared annotation noise and observation bias may affect validation reliability and the overall model evaluation.

Response from Authors: We appreciate this suggestion regarding validation independence. We would like to clarify the composition of GlobalGeoTree to address the concern about "volunteer-contributed" bias.

Hybrid Data Composition: GlobalGeoTree is not solely derived from opportunistic crowdsourcing. It is a hybrid integration of:

- **High-Quality Citizen Science (∼70%):** Sourced exclusively from "Research Grade" iNaturalist observations, which require community consensus for identification.

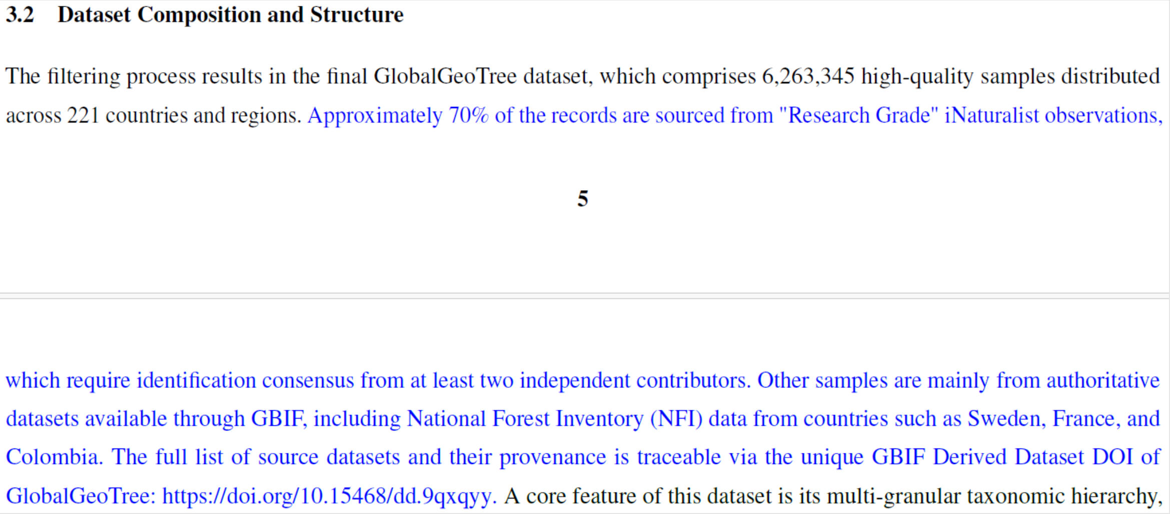
- **Official Inventories & NFIs:** We have proactively integrated available authoritative datasets, including National Forest Inventories (e.g., from Sweden, France, Colombia) and academic survey plots.

Full provenance for all constituent datasets is transparently tracked via our GBIF Derived Dataset

DOI: `https://www.gbif.org/derivedDataset/10.15468/dd.9qxqyy`.

Because we aimed to maximize the model's global generalization capability, we chose to assimilate these high-quality official datasets into the training distribution rather than holding them out as a small external validation set. Excluding them from training solely for validation would have degraded the model's performance in those key regions.

We have updated **Section 3.2** to explicitly state the proportion of Research Grade data and the inclusion of NFI sources. We also added a discussion in **Section 7.2 (Limitations)** acknowledging that shared observation biases still exist.

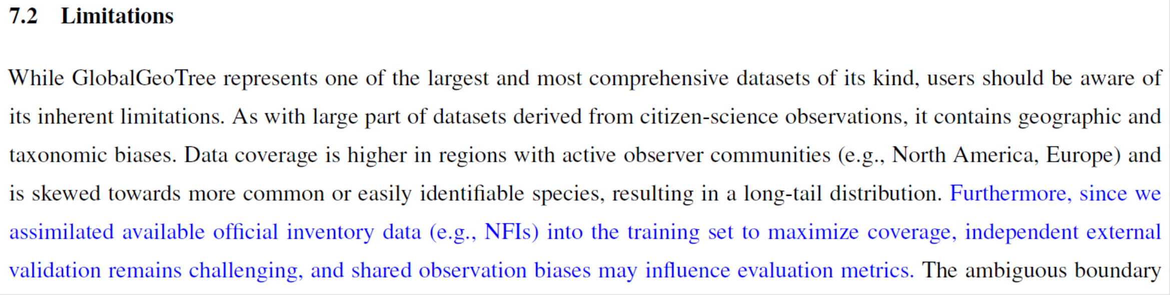


3.2 Dataset Composition and Structure

The filtering process results in the final GlobalGeoTree dataset, which comprises 6,263,345 high-quality samples distributed across 221 countries and regions. Approximately 70% of the records are sourced from "Research Grade" iNaturalist observations,

5

which require identification consensus from at least two independent contributors. Other samples are mainly from authoritative datasets available through GBIF, including National Forest Inventory (NFI) data from countries such as Sweden, France, and Colombia. The full list of source datasets and their provenance is traceable via the unique GBIF Derived Dataset DOI of GlobalGeoTree: https://doi.org/10.15468/dd.9qxqyy. A core feature of this dataset is its multi-granular taxonomic hierarchy,



7.2 Limitations

While GlobalGeoTree represents one of the largest and most comprehensive datasets of its kind, users should be aware of its inherent limitations. As with large part of datasets derived from citizen-science observations, it contains geographic and taxonomic biases. Data coverage is higher in regions with active observer communities (e.g., North America, Europe) and is skewed towards more common or easily identifiable species, resulting in a long-tail distribution. Furthermore, since we assimilated available official inventory data (e.g., NFIs) into the training set to maximize coverage, independent external validation remains challenging, and shared observation biases may influence evaluation metrics. The ambiguous boundary

3. Structured Summary Table

The manuscript states that 21,001 species are included. I suggest adding a structured summary table that includes representative families, genera, and species; their geographic distribution or dominant regions; brief ecological descriptions; and, optionally, sample images from public datasets. This would greatly enhance the readability and interpretability of the dataset.

Response from Authors: We agree that visualizing the taxonomic breadth is vital for readability. We have addressed this through two complementary approaches:

1. **Representative Species Table:** We have added a new table in Appendix A (Table A1) that profiles representative species from major biomes (e.g., *Picea abies* for Boreal, *Eucalyptus* for Australasia, *Quercus* for Temperate). For each entry, we detail the family, genus, dominant region, and key ecological traits, providing a structured "snapshot" of the dataset's diversity as requested.

2. **Interactive Exploration (GEE App):** To maximize interpretability beyond static tables, we developed the **GlobalGeoTree Explorer App** (see Fig. 5 in the revised manuscript). This tool allows users to dynamically query any individual sample, view its geographic distribution on a satellite map, and inspect detailed metadata interactively. The explorer is available at: `https://ee-yangm.projects.earthengine.app/view/globalgeotree-explorer`.

We have inserted the representative species table into **Appendix A** and prominently referenced the GEE App in **Section 4.1** as a tool for dataset exploration.

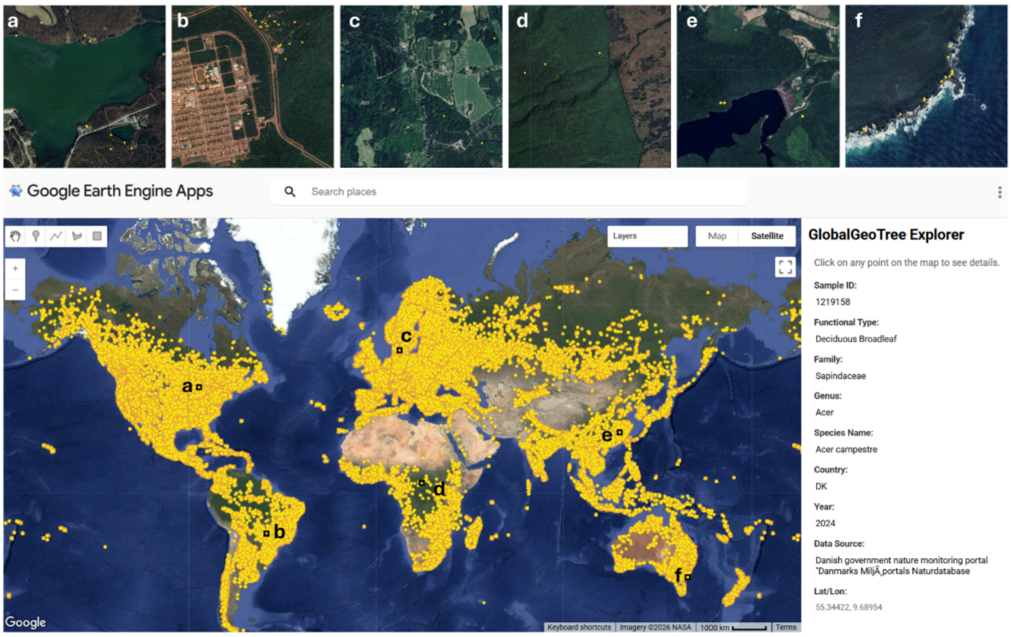
A5 Representative Species Profile

Table A1 provides a structured summary of representative species from the GlobalGeoTree dataset, highlighting the diversity across different functional types and geographic regions.

Table A1. Structured summary of representative tree species in GlobalGeoTree, spanning major functional types and biogeographical realms.

Functional Type	Family	Genus	Species	Dominant Region	Ecological Context
Evergreen Needleleaf	Pinaceae	*Picea*	*Picea abies*	Europe	Dominant boreal forest conifer
Evergreen Needleleaf	Pinaceae	*Pseudotsuga*	*Pseudotsuga menziesii*	North America	Key timber species of Pacific Northwest
Deciduous Broadleaf	Fagaceae	*Quercus*	*Quercus robur*	Europe	Keystone species of temperate woodlands
Deciduous Broadleaf	Sapindaceae	*Acer*	*Acer saccharum*	North America	Major component of northern hardwood forests
Evergreen Broadleaf	Myrtaceae	*Eucalyptus*	*Eucalyptus globulus*	Oceania	Fast-growing, widely distributed in Australia
Evergreen Broadleaf	Urticaceae	*Cecropia*	*Cecropia peltata*	South America	Pioneer species in neotropical rainforests
Deciduous Needleleaf	Pinaceae	*Larix*	*Larix decidua*	Europe	High-altitude conifer of the Alps

4 Data Quality and Validation



GlobalGeoTree Explorer

Click on any point on the map to see details.

Sample ID:
1219158

Functional Type:
Deciduous Broadleaf

Family:
Sapindaceae

Genus:
Acer

Species Name:
Acer campestre

Country:
DK

Year:
2024

Data Source:
Danish government nature monitoring portal
"Danmarks MiljÃ¸portals Naturdatabase

Lat/Lon:
55.34422, 9.68954

Figure 5. The GlobalGeoTree Explorer App. The interface allows users to visualize the global distribution of tree samples (bottom map) and inspect individual points against high-resolution satellite imagery (top panels a-f). Detailed attributes for selected samples are displayed in the sidebar, facilitating transparent community validation.

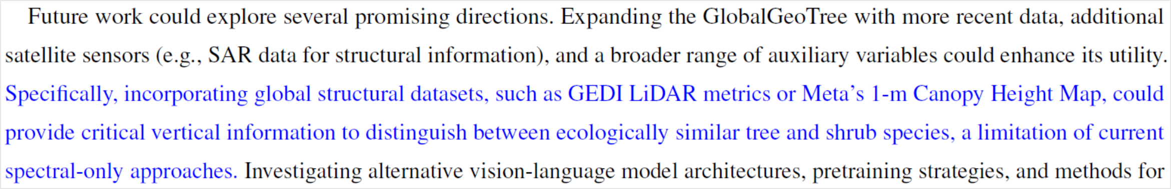
4. Structural Canopy Information

The current framework relies on multispectral imagery and environmental variables but lacks structural canopy information, which is highly relevant for separating woody species. I suggest simply discussing the potential benefits of incorporating global 1-m canopy height maps (e.g., Meta's global CHM), global-scale LiDAR-derived products, or other structural/vertical metrics.

Response from Authors: We fully agree that structural information is a critical missing modality, particularly for distinguishing between tree and shrub forms within the same genus or separating mature forests from young plantations.

While our current framework prioritizes the global consistency and temporal richness of Sentinel-2 time series, we recognize that integrating vertical structure metrics, such as Global Ecosystem Dynamics Investigation (GEDI) waveforms or Meta's global CHM, would enhance discriminative power.

We have expanded the **Section 9 (Future Work)** to explicitly discuss the integration of global structural products.



Future work could explore several promising directions. Expanding the GlobalGeoTree with more recent data, additional satellite sensors (e.g., SAR data for structural information), and a broader range of auxiliary variables could enhance its utility. Specifically, incorporating global structural datasets, such as GEDI LiDAR metrics or Meta's 1-m Canopy Height Map, could provide critical vertical information to distinguish between ecologically similar tree and shrub species, a limitation of current spectral-only approaches. Investigating alternative vision-language model architectures, pretraining strategies, and methods for

5. Global Land-Cover Products

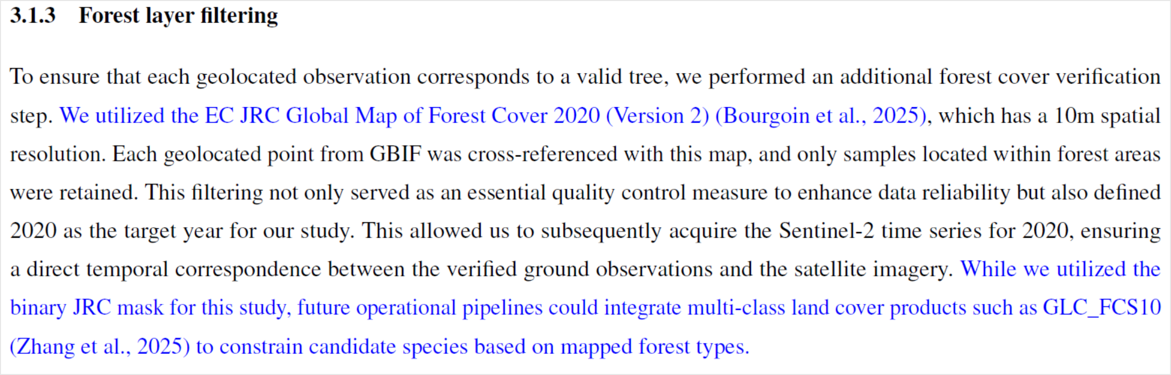
Products such as GLC_FCS10 provide 10-m vegetation functional-type information and could be valuable for excluding non-forested regions or constraining the candidate species space. A short discussion on the potential integration of global land-cover products would be beneficial.

Response from Authors: We appreciate this suggestion. While we currently utilize the **JRC Global Map of Forest Cover 2020** (10m) to mask non-forested regions (Section 3.1.3), we agree that finer-grained products like GLC_FCS10 offers additional utility.

Integrating such products could refine the inference pipeline by acting as a **"Semantic Constraint Mask."** For instance, if GLC_FCS10 classifies a pixel as "Needleleaf Forest," the model's probability space could be constrained to only needleleaf species, thereby reducing false positives from spectrally similar broadleaf species.

We have implemented a similar strategy in our new **Data Quality and Validation** section. We utilized the *Copernicus Global Land Cover Layers* (CGLS-LC100), which provides discrete forest functional types similar to GLC_FCS10, to perform a global cross-validation. Our analysis (Fig. 6 in the revised manuscript) revealed an **80.38% overall agreement** between our dataset's leaf-type labels and the CGLS product globally. This high consistency confirms that such land-cover maps can indeed serve as effective semantic constraints.

We have added a short discussion in **Section 3.1.3** and **Section 4.2** highlighting this synergy, suggesting products like GLC_FCS10 for operational constraints.



3.1.3 Forest layer filtering

To ensure that each geolocated observation corresponds to a valid tree, we performed an additional forest cover verification step. We utilized the EC JRC Global Map of Forest Cover 2020 (Version 2) (Bourgoin et al., 2025), which has a 10m spatial resolution. Each geolocated point from GBIF was cross-referenced with this map, and only samples located within forest areas were retained. This filtering not only served as an essential quality control measure to enhance data reliability but also defined 2020 as the target year for our study. This allowed us to subsequently acquire the Sentinel-2 time series for 2020, ensuring a direct temporal correspondence between the verified ground observations and the satellite imagery. While we utilized the binary JRC mask for this study, future operational pipelines could integrate multi-class land cover products such as GLC_FCS10 (Zhang et al., 2025) to constrain candidate species based on mapped forest types.

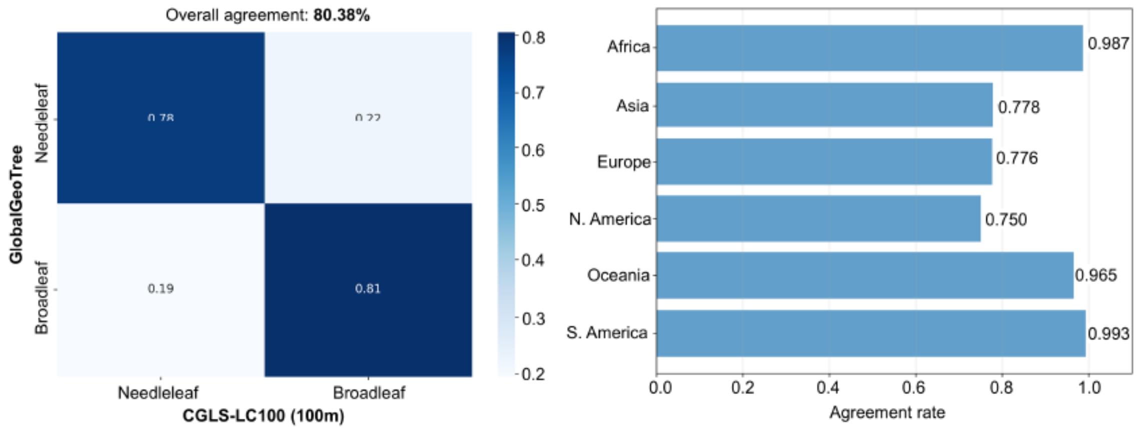


Figure 6. Cross-validation results against the CGLS-LC100 product. **Left:** Normalized confusion matrix at the Leaf Level (Needleleaf vs. Broadleaf), showing an overall agreement of 80.38%. **Right:** Agreement rates broken down by continent, highlighting strong consistency in tropical regions.

6. Real-World Inference and Operational Mapping

The proposed dataset is highly valuable for global-scale tree species mapping. However, the manuscript does not sufficiently address how the dataset can be used in real-world inference and operational mapping. I recommend simply adding a dedicated section that discusses practical workflows for deploying the dataset in large-scale tree species mapping, potential inference pipelines, and challenges in global deployment (e.g., domain shifts, spatial biases, and seasonal variability). Such a discussion would help bridge the gap between dataset creation and applied remote-sensing or ecological use cases.

Response from Authors: We agree that outlining practical deployment strategies is crucial for maximizing the dataset's utility. We have revised and expanded a new section **Potential Impact and Operational Deployment (Section 7.3)** to explicitly discuss operational workflows.

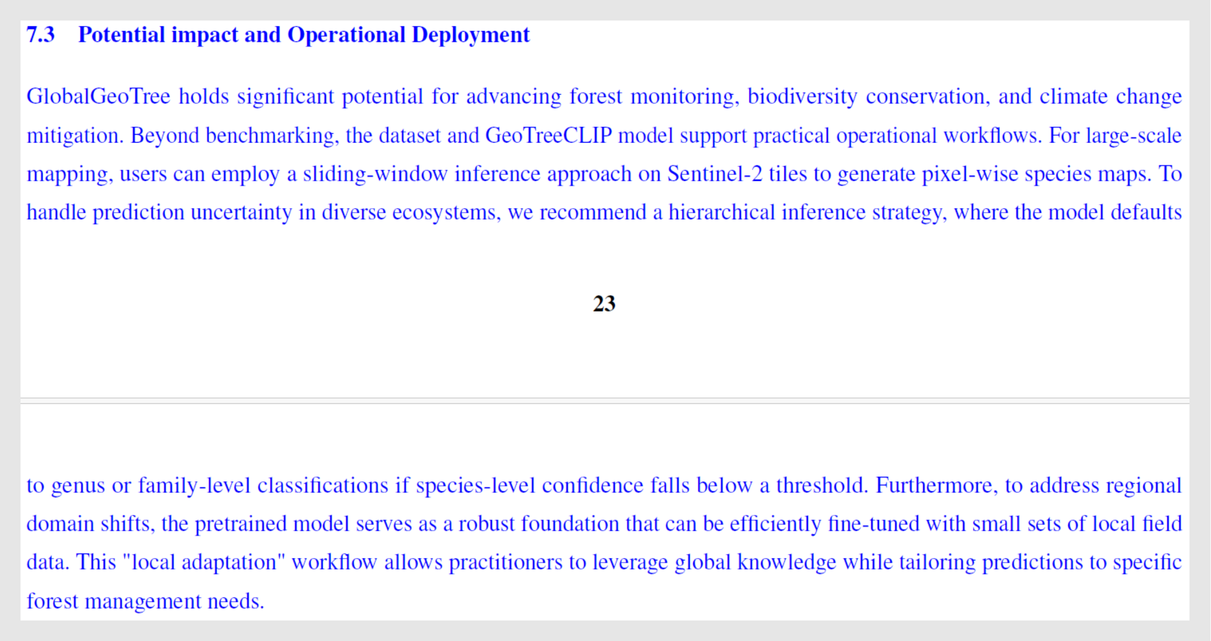
This section outlines a practical workflow for large-scale mapping:

1. **Inference Pipeline:** We propose a sliding-window approach where the pretrained GeoTreeCLIP model processes Sentinel-2 tiles to generate pixel-wise species probability maps.

2. **Hierarchical Prediction:** To handle uncertainty, we suggest a hierarchical inference strategy, outputting high-confidence species labels where possible, but defaulting to genus or family levels when model confidence is low.

3. **Addressing Domain Shifts:** We recommend a "Local Few-Shot Adaptation" workflow, where users fine-tune the global model with a small set of local field plots before deploying it to a specific region, thereby mitigating spatial biases.

We have revised and expanded **Section 7.3** to integrate these practical deployment strategies, bridging the gap between the benchmark dataset and real-world forest monitoring applications.



7.3 Potential impact and Operational Deployment

GlobalGeoTree holds significant potential for advancing forest monitoring, biodiversity conservation, and climate change mitigation. Beyond benchmarking, the dataset and GeoTreeCLIP model support practical operational workflows. For large-scale mapping, users can employ a sliding-window inference approach on Sentinel-2 tiles to generate pixel-wise species maps. To handle prediction uncertainty in diverse ecosystems, we recommend a hierarchical inference strategy, where the model defaults to genus or family-level classifications if species-level confidence falls below a threshold. Furthermore, to address regional domain shifts, the pretrained model serves as a robust foundation that can be efficiently fine-tuned with small sets of local field data. This "local adaptation" workflow allows practitioners to leverage global knowledge while tailoring predictions to specific forest management needs.