TREEUQ: UNCERTAINTY-AWARE MULTIMODAL REGRESSION FOR TREE STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate characterization of forest structure is essential for understanding the global carbon cycle and ecosystem dynamics. While high resolution aerial imagery has enabled large-scale mapping of continuous forest attributes like tree segmentation, fundamental ecological quantities such as tree height and density remain largely absent from open satellite resolution (Sentinel-1 and Sentinel-2) benchmarks. In this work, we introduce TreeUQ, a multimodal Earth observation benchmark for tree-structure estimation over Bavaria, Germany. The dataset contains Sentinel-1 and Sentinel-2 composites, high resolution RGB (20cm), tree-species information, and six tree-structure related targets. We benchmark mean tree-height and density estimation, and compare methods for uncertainty quantification. Modality ablations assess the contributions of radar, optical, and species information. We establish rigorous evaluation protocols using a spatially blocked split strategy to prevent geographic data leakage.

1 INTRODUCTION

Accurate, large-scale characterization of forest structure is essential for understanding the global carbon cycle, biodiversity, and ecosystem resilience. Recent advances in satellite remote sensing have enabled the development of high-resolution (10 m) maps of forest attributes such as canopy height, biomass, and land cover across continental scales (see Potapov et al. (2021) Lang et al. (2023)). These datasets have become critical inputs to downstream applications in climate science, ecology, and environmental monitoring.

However, a critical limitation of existing datasets and benchmarks is that they focus almost exclusively on point estimates of forest attributes, without accounting for the inherent uncertainty in these predictions. In real-world applications, uncertainty is not only a secondary consideration. It directly affects decision-making in domains such as carbon accounting, forest management, and risk assessment (see Kujala et al. (2023)). At the same time, existing benchmarks focus almost exclusively on continuous structural variables (e.g., canopy height or biomass) or categorical outputs (e.g., forest type or species) (see Su et al. (2025) , Jiang & Neumann (2025)). A fundamental ecological quantity, namely tree density, defined as the number of trees per unit area, remains absent from large-scale, open benchmarking datasets derived from medium-resolution satellite imagery (see Wang et al. (2025)). The reason for this gap is rather obvious: at a spatial resolution of 10 m, individual tree crowns are typically unresolved or at least difficult to detect, making direct counting infeasible and requiring models to infer tree density indirectly from spectral, temporal, and structural signals. While our ground-truth inventory is geographically focused on the diverse biomes of Bavaria, our inputs (Sentinel-1 and Sentinel-2) are globally available. Consequently, models trained on this dataset can be deployed for inference, providing a critical foundational step toward planetary-scale tree density mapping.

In this work, we introduce TreeUQ, a large-scale, multimodal dataset designed to explicitly study uncertainty in the Earth observation domain. By fusing tree species (per pixel), multi-seasonal Sentinel-1 and Sentinel-2 imagery, as well as high-resolution RGB orthophotos with a comprehensive individual-tree inventory across the federal state of Bavaria, a region characterized by diverse forest type and topographic conditions, we provide a challenging and representative base for testing machine learning models. We position this dataset as a benchmark for multiple target learning in Earth observation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

More precisely, the novel contributions of our paper are threefold: (1) We introduce a dataset with estimation of tree counts and average tree height at 10 m resolution from Sentinel-1 and Sentinel-2 data as well as high-resolution RGB orthophotos, covering the whole of Bavaria where the ground-truth is derived from a comprehensive individual-tree inventory. (2) The dataset includes properties such as variance of the tree heights as well as density of the tree counts by taking into account the spatial structure of the data (more precisely described in Section 3). (3) We describe a training/evaluation protocol, as well as baseline models establishing reference performance for future work.

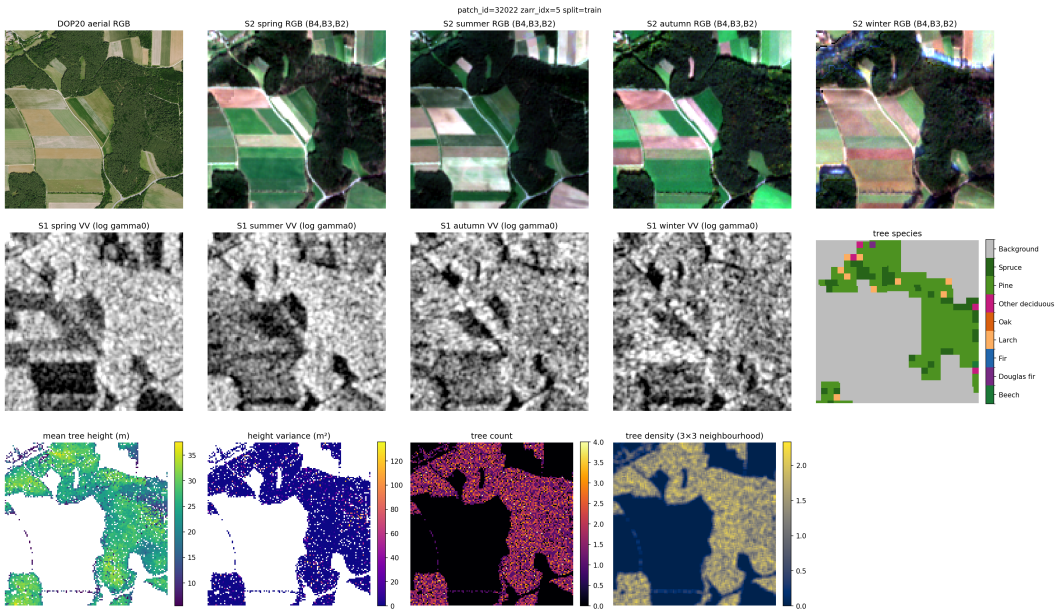

Figure 1: Overview of TreeUQ dataset. The dataset provides multimodal inputs, including Sentinel-2 MSI, Sentinel-1 SAR, and categorical tree species maps. Models must utilize these inputs to predict continuous mean tree height and tree density. Additionally, high-resolution 20cm RGB orthophotos are available for the region of Bavaria.

2 Related Work

2.1 Forest structure mapping from satellite imagery

Spaceborne GEDI LiDAR and dense satellite imagery have enabled forest-structure mapping from regional to global scales. (Kacic et al., 2023) combine GEDI supervision with Sentinel-1 and Sentinel-2 inputs in a random forest to estimate canopy height, canopy cover, and above-ground biomass across Germany at 10 m resolution. Deep models now dominate this setting: (Lang et al., 2023) use a probabilistic fully convolutional network with GEDI and Sentinel-2 to produce a global 10 m canopy-height map, while (Su et al., 2025) and (Schwartz et al., 2024) use U-Net-based models for high-resolution regional products. (Morin et al., 2024) further scale GEDI-supervised height and biomass mapping to national coverage.

Work that explicitly models uncertainty remains more limited. (Lang et al., 2022) regress global canopy-top height directly from GEDI LiDAR waveforms with an ensemble of convolutional networks, obtaining both a point prediction and predictive uncertainty at each GEDI footprint. More recently, (Bader et al., 2026) adapt deep evidential regression to sparse multimodal Sentinel-1/-2 supervision. Their masked objective predicts aleatoric and epistemic uncertainty in one forward pass. These approaches motivate uncertainty-aware evaluation, but neither provides a benchmark that jointly represents continuous height, density, and supporting within-cell statistics.

Related datasets address other forest attributes: (Ahlswede et al., 2023) and (Blickensdörfer et al., 2024) study species mapping from aerial or multitemporal Sentinel data. Overall, existing medium-

resolution products still emphasize continuous variables such as height, biomass, and canopy cover rather than tree density or joint continuous–count prediction.

2.2 Tree detection and counting

Individual-tree detection and counting are well studied when crowns are directly visible in aerial, UAV, or airborne-LiDAR data, commonly through object detection or instance segmentation (see Tong & Zhang (2025), Weishaupt et al. (2025), Freudenberg et al. (2022), Yao et al. (2021)). Such data are costly and spatially limited, even large annotated collections may cover only about 100 hectares (Troles et al., 2024).

At medium resolution, (Brandt & Stolle, 2021) show that Sentinel optical and radar imagery can identify trees with canopy diameters above three meters, but do not estimate per-pixel density or other tree attributes. Bavarian governmental products likewise provide automatically derived tree counts (Vermessungsverwaltung, 2025), but are not aligned with Sentinel prediction tasks and do not supply standardized machine-learning splits or evaluation protocols.

2.3 Multimodal and multi-task learning in remote sensing

Multimodal fusion combines complementary SAR sensitivity to canopy structure and moisture with optical information about vegetation condition and phenology (Li et al., 2022b;a). Proposed architectures include dual-stream networks, attention-based fusion, and transformers (Roy et al., 2023; Quan et al., 2024); in forestry, Sentinel-1/-2 and GEDI combinations have been effective for canopy-height and biomass mapping (Kacic et al., 2023; Schwartz et al., 2024).

Multi-task remote-sensing models typically share an encoder while retaining task-specific heads. Prior work jointly predicts closely related outputs such as height and semantic labels (Lu et al., 2022; Li et al., 2025). TreeUQ instead couples continuous height regression with density estimation and uncertainty assessment, offering a distinct test of shared representations across target types.

2.4 Positioning of our work

Prior work therefore either detects individual trees in high-resolution imagery or predicts continuous forest attributes at medium resolution. To the best of our knowledge, TreeUQ is the first 10 m Sentinel benchmark to combine tree height, tree density, and supporting within-cell statistics. It links large-scale forest mapping with multi-task and uncertainty-aware learning without assuming that individual crowns are resolved.

3 The TreeUQ Dataset

We introduce a large-scale, multimodal Earth observation dataset for estimation of tree density and tree height statistics at 10 m resolution. The dataset is constructed over Bavaria and combines satellite observations from Sentinel-1, Sentinel-2 and tree species labels with individual-tree inventory data. It supports multi-task regression, tree-species segmentation, multimodal fusion, spatial-generalization evaluation, and uncertainty quantification. The dataset is defined on a common 10 m grid (EPSG:25832) covering approximately 367×358 km, corresponding to $36,709 \times 35,856$ pixels (also referred to as cells). All inputs and labels are spatially aligned to this reference grid. The dataset integrates the following primary data sources aligned on a common 10 m resolution grid:

1. **Sentinel-2 multispectral imagery**(European Space Agency, 2026b): Level-2A scenes from 2025 with scene cloud cover below 20% are masked using the Scene Classification Layer (clouds, cloud shadow, cirrus, snow, and saturated pixels) and aggregated into four seasonal per-pixel median composites, from spring 2025 to winter 2025/26. Each composite contains ten bands (B2-B8, B8A, B11, B12). Moreover, the 20 m bands are bilinearly resampled to 10 m.
2. **Sentinel-1 SAR imagery**(European Space Agency, 2026a): Radar observations (VV, VH polarization) are similarly aggregated into seasonal composites. As an active sensing modality, SAR provides structural information that is robust to cloud cover and illumination, complementing optical features.

3. **DOP20 aerial imagery**(Landesamt für Digitalisierung, 2025a): 20 cm RGB orthophotos from the two-year Bavarian flight cycle provide crown-level texture beyond satellite resolution. A chip is released when at least 50% of its footprint is covered, which applies to 96.5% of all patches.
4. **Tree species raster**(Mu et al., 2025): A 10 m map of eight dominant species (beech, oak, other deciduous, larch, spruce, pine, fir, Douglas fir). We reproject it to the reference grid with nearest-neighbour resampling.
5. **Individual-tree inventory**Landesamt für Digitalisierung, 2025b: A large-scale dataset of geolocated tree points with associated height attributes. In addition to defining supervision labels, it determines the spatial extent and alignment of the dataset.

Supervision is derived from an individual-tree inventory dataset (Vermessungsverwaltung, 2025) containing geolocated tree positions and associated tree height measurements. Individual-tree observations are rasterized onto the 10 m grid without buffering tree crowns or spreading individual records across neighboring cells. Every tree is assigned to the grid cell containing its point location. For each cell containing one or more observations, we measure the number of trees, mean tree height, median tree height, and variance of tree height. Cells without inventory observations contain zero for `tree_count` and missing values for the height targets. Two additional measures describe the local distribution of tree counts: `tree_density` is the mean tree count within a 3×3-pixel neighborhood, corresponding to a 30×30 m context window, and `tree_count_variance` is the spatial variance of tree counts within the same neighborhood.

Across the complete rectangular grid, 17.86% of all cells contain at least one tree observation. The dataset is organized into non-overlapping 128×128-pixel patches, each covering an area of 1.28×1.28 km. Every patch contains four Sentinel-2 tensors of shape $128 \times 128 \times 10$, four Sentinel-1 tensors of shape $128 \times 128 \times 2$, one categorical species raster, and six 128×128 tree-structure target maps. For 96.5% of all patches, a co-registered $6{,}400 \times 6{,}400 \times 3$ DOP20 chip covers the same geographic footprint. Patches in which fewer than 50% of Sentinel-2 pixels are valid after cloud and snow masking are discarded. No gap-filling is applied. The release is distributed through 837 geographic tiles, each associated with a fixed 10×10 km indexing cell. The core component (Sentinel-1, Sentinel-2, species, and labels) occupies 151 GiB and is distributed separately from the 5.1 TiB DOP20 collection, so users can retrieve it without downloading the aerial imagery.

To reduce evaluation bias from spatial autocorrelation, we use geographic block splitting. The dataset is divided into a 6×6 grid of blocks,(i.e. approximately 60×61 km), and entire blocks are assigned to the training, validation, or test partition. Validation and test blocks are selected at random to provide geographically distributed holdout regions. After filtering to Bavaria region, this yields 45,419 patches: 31,806 training, 7,048 validation, and 6,565 test patches. Partitions are adjacent only along block boundaries, where some residual spatial autocorrelation between training and evaluation patches may remain. A detailed description of all data sources and processing steps is given in Appendix A.1.A.2.

4 Benchmarking methods and protocol

4.1 Task Formulation: Height And Density Estimation

The primary downstream task established by this benchmark is pixel-wise regression at 10 m spatial resolution. Unlike traditional high-resolution object detection tasks that draw bounding boxes around individual crowns, this task requires models to infer aggregated, sub-pixel properties directly from medium-resolution radiometric and structural signals.

The primary benchmark considers two complementary settings:

1. **Point estimation**, in which the model predicts a single expected tree height and tree density for every valid target pixel.
2. **Uncertainty quantification**, in which the model additionally estimates a predictive distribution or calibrated prediction interval around each height and density estimate.

Mean-height targets are missing in cells without inventory trees. We therefore define a patch-wise binary supervision mask $M \in \{0,1\}^{128\times128}$ where $M_{ij} = 1$ if the pixel at position (i, j) within the

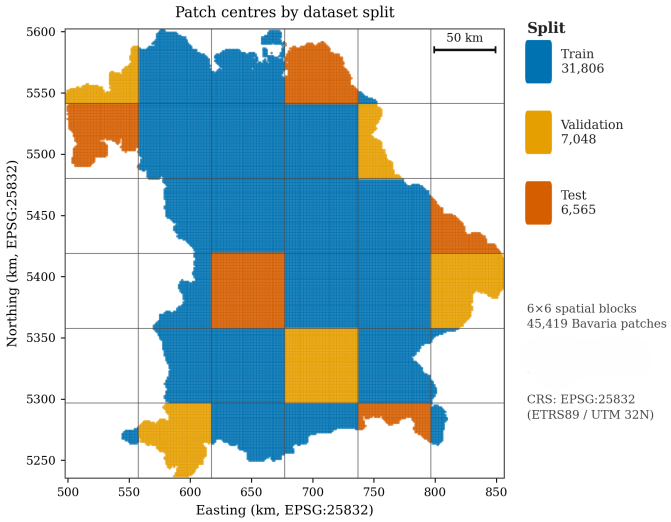

Figure 2: Spatial block splitting strategy across Bavaria. To prevent geographic data leakage, the region is divided into a 6×6 grid. Patches are strictly assigned to Train (blue), Validation (orange), or Test (red) sets based on their macro-block.

patch has a valid tree height label. For the computation of training losses and evaluation metrics, we only include pixels for which $M_{ij} = 1$. Consequently, missing height annotations are not interpreted as trees of zero height. The released targets remain in meters, while a $\log(1 + y)$ transformation is applied within the modeling pipeline (see Figure 3). Predictions are mapped back to meters using a $\exp(\hat{y}) - 1$ transformation before evaluation.

For density, the target is the mean tree count in the 3×3 neighbourhood of each cell. All cells are passed to the model as input, but training losses and evaluation metrics are computed only on cells with density greater than zero.

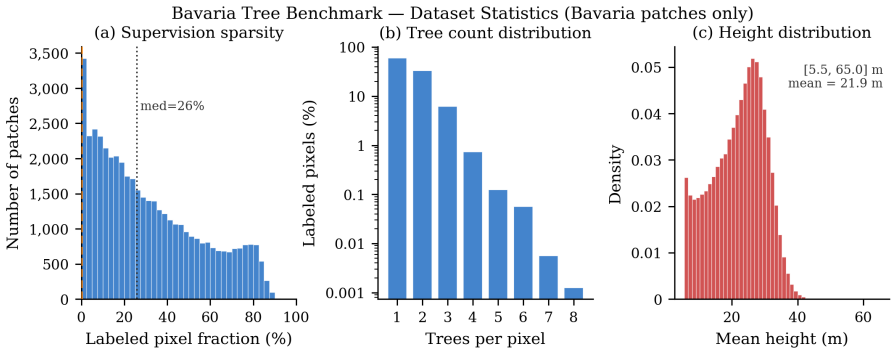

Figure 3: Dataset statistics and label distributions. (a) The dataset exhibits spatial sparsity, with a median valid label coverage of 26% per patch. (b) Tree counts per 10 m×10 m pixel follow a heavy-tailed distribution. (c) Mean tree heights range from 5.5 m to over 60 m. Due to these skewed distributions, target variables are $\log(1 + y)$ transformed during training.

4.2 Data Preprocessing for Modeling

To frame the problem for machine learning models, the varied input modalities are structurally standardized. Rather than using explicit temporal embeddings (for models other than Prithvi), seasonality is implicitly encoded by stacking the temporal composites along the channel dimension. The standard input tensor for each 128×128 spatial patch consists of 49 channels: 8 channels of Sentinel-1 data (4 seasons×2 bands), 40 channels of Sentinel-2 data (4 seasons×10 bands), and 1 categorical tree species channel. For the target variables, the previously mentioned $\log(1 + y)$-transformation is applied during preprocessing to stabilize the variance of the gradients during training.

4.3 Point-Estimation Baselines

We evaluate five approaches spanning non-spatial machine learning, conventional convolutional networks, dense vision transformers, and Earth observation foundation models:

- **XGBoost** (Chen & Guestrin, 2016): XGBoost provides a non-spatial baseline that predicts each pixel independently.
- **U-Net** (Ronneberger et al., 2015): A U-Net encoder-decoder provides a convolutional dense-regression baseline.
- **SegFormer-B3** (Xie et al., 2021): SegFormer-B3 serves as the principal transformer-based dense prediction architecture.
- **Prithvi** (Szwarcman et al., 2026): Prithvi is an Earth observation foundation model pre-trained on multispectral satellite time series.
- **DOFA** (Xiong et al., 2025): DOFA is a multimodal Earth observation foundation model designed to accommodate sensors with different spectral configurations through wavelength-aware dynamic encoding.

4.4 Uncertainty-Quantification Methods

All uncertainty experiments use SegFormer-B3 as the common backbone, ensuring that differences in uncertainty quantification quality arise primarily from the uncertainty quantification method rather than model capacity. We evaluate four approaches:

- **Deep ensembles** (Rahaman & Thiery, 2021): We train 3 SegFormer-B3 models independently using different parameter initializations and data-ordering seeds.
- **Monte Carlo dropout** (Gal & Ghahramani, 2016): Dropout layers remain active during inference, and each test patch is evaluated 20 times.
- **Deep evidential regression** (Amini et al., 2020): The network predicts the parameters of a Normal–Inverse-Gamma distribution rather than a single height or density value.
- **Conformalized quantile regression** (Rossellini et al., 2024): SegFormer-B3 is trained to predict lower and upper conditional height and density quantiles using the pinball loss. The resulting intervals are calibrated using a held-out calibration subset from the validation data.

4.5 DOP20-Fused Height and Density Prediction

We evaluate whether 20 cm DOP20 imagery improves height and density prediction. Because each patch contains a $6{,}400 \times 6{,}400$ RGB image, we train on a fixed 10% subset of the training partition and evaluate on the full DOP20-covered validation and test splits. We compare three single-task U-Net settings with ImageNet-pretrained ResNet50 encoders: (i) *S1+S2*, (ii) *S1+S2+DOP20*, which adds precomputed DOP20 features at the decoder skip connections and (iii) *DOP20 RGB*, which uses only orthophotos area-averaged to 10 m. DOP20 features are obtained by encoding the 50×50-pixel orthophoto tile of each 10 m cell with a frozen ImageNet ResNet50 and reducing the 2048-dimensional embeddings to 256 channels with PCA fitted on training patches only. DOP20 inputs are available at test time in all settings (Appendix E.3).

4.6 Evaluation Metrics

All metrics are computed on the original scale over test pixels with valid supervision. Point estimates are evaluated with RMSE and MAE. Uncertainty is evaluated at a nominal 90% level using empirical interval coverage (Cov.) and mean prediction-interval width (MPIW). We further report the expected normalized calibration error (ENCE) (Levi et al., 2022), which compares predicted and observed error across bins of predicted uncertainty, and the continuous ranked probability score (CRPS) (Gneiting & Raftery, 2007), which jointly rewards calibration and sharpness of the predictive distribution.

4.7 Experimental setting

All neural models are implemented in PyTorch 2.1.2 and trained using three fixed random seeds. We report the mean and standard deviation across runs. Training augmentation consists of random horizontal and vertical flips ($p = 0.5$ each) and rotations by multiples of 90° ($p = 0.75$). Each geometric transformation is applied identically to the input, target, and validity mask. Hyperparameters and stopping criteria are selected exclusively on the validation split, and the geographically held-out test set is evaluated only after model selection. Input normalization, target transformation, masked optimization, and evaluation metrics follow the protocol described above.

5 Results and discussion

The evaluation of our baseline models highlights the significant challenge posed by the TreeUQ benchmark, particularly in estimating heterogeneous forest structure targets under label sparsity. The results across three random seeds are summarized in the tables below.

5.1 Tree Height and Density Estimation

For Tree height prediction, the non-spatial XGBoost baseline still reaches an RMSE of 6.870 ± 0.002 m and an MAE of 5.487 ± 0.003 m, confirming that raw spectral and temporal signatures contain useful height information. U-Net improves on that baseline, with an RMSE of 6.301 ± 0.026 m and an MAE of 4.941 ± 0.016 m, showing that spatial context matters for continuous canopy structure. SegFormer-B3 is essentially level with U-Net (RMSE 6.312 ± 0.110 m, MAE 4.923 ± 0.103 m), but the difference sits inside the transformer's larger spread. Fine-tuned DOFA lowers error further, to an RMSE of 6.24 m and an MAE of 4.86 m. The highest overall performance on the height task is Prithvi 2.0, at an RMSE of $\mathbf{5.37}$ m and an MAE of $\mathbf{4.06}$ m, indicating that a large geospatial foundation model pretrained on Earth observation data transfers more effectively to this sparse inventory-height regression task.

For tree density, XGBoost attains the lowest RMSE (0.344 ± 0.000 trees per 10 m cell) with an MAE of 0.248. SegFormer-B3 and U-Net reach a marginally lower MAE (0.244 ± 0.005 and 0.244 ± 0.002), a difference within seed variation, and SegFormer-B3 stays within 0.004 RMSE of XGBoost (0.348 ± 0.005). U-Net's higher and less stable RMSE (0.391 ± 0.059) [is driven by one of its three seeds]. Prithvi-EO-2.0, which led on height, remains competitive (RMSE 0.352 ± 0.001, MAE 0.249 ± 0.000), whereas DOFA has the highest MAE (0.277 ± 0.001, RMSE 0.377 ± 0.001). Unlike for height, neither spatial context nor foundation-model pretraining yields a clear gain over pixel-wise features for density at this resolution. Modality ablation (Appendix 5) further indicates that, under our preprocessing, Sentinel-2 carries most of usable height signal, while adding Sentinel-1 amplitude or the species channel yields no measurable gain.

Table 1: Point estimates on the original scale. Height in meters, density is trees per 10 m cell.

	Height (m)		Density	
Model	RMSE	MAE	RMSE	MAE
XGBoost	6.870 ± 0.002	5.487 ± 0.003	$\mathbf{0.344 \pm 0.000}$	0.248 ± 0.000
U-Net	6.301 ± 0.026	4.941 ± 0.016	0.391 ± 0.059	0.244 ± 0.002
SegFormer-B3	6.312 ± 0.110	4.923 ± 0.103	0.348 ± 0.005	$\mathbf{0.244 \pm 0.005}$
DOFA ViT-Large	6.24 ± 0.030	4.86 ± 0.010	0.377 ± 0.001	0.277 ± 0.001
Prithvi 2.0	$\mathbf{5.37 \pm 0.020}$	$\mathbf{4.06 \pm 0.012}$	0.352 ± 0.001	0.249 ± 0.000

5.2 High-Resolution Tree Height and Density Estimation

On the 10% height split, the model trained on Sentinel-1 and Sentinel-2 data only reaches an RMSE of 6.65 m, an MAE of 5.28 m, and an R^2 of 0.331. Fusing those inputs with frozen DOP20 features (ResNet50/PCA maps) is the strongest setting, with an RMSE of $\mathbf{5.64}$ m, an MAE of $\mathbf{4.38}$ m, and an R^2 of $\mathbf{0.519}$. The aerial structural prior transfers even when inventory labels are sparse. Native 20 cm RGB is weaker than the satellite (S1, S2) baseline with RMSE 7.20 m, MAE 5.81 m and R^2 0.217. High-resolution therefore helps height when it is encoded as frozen features and combined with seasonal S1+S2, while raw downsampled RGB does not replace that multisensor stack. Density on the same 10% split shows a larger relative gain from the same fusion. S1+S2 alone yields an RMSE of 0.352, an MAE of 0.256, and an R^2 of 0.299. Privileged DOP20 again leads, at an RMSE of $\mathbf{0.274}$, an MAE of $\mathbf{0.201}$, and an R^2 of $\mathbf{0.574}$ — a drop of 0.078 RMSE and 0.055 MAE, and a rise of 0.275 in R^2. DOP20 RGB already improves the satellite baseline (RMSE 0.340, MAE 0.248, R^2 0.343), which fits a count task that can use fine canopy texture, but it remains well behind the fused model. Across both targets, frozen aerial features add information that neither seasonal radar–optical inputs nor downsampled RGB capture on their own, with the clearest lift on density.

Table 2: Height and density trained on a 10% subset of the training split. S1+S2+DOP20 fuses frozen ResNet50 features of 20 cm orthophotos (PCA-reduced to 256 channels) with seasonal S1+S2

	Height			Density		
Setting	RMSE (m)	MAE (m)	R^2	RMSE	MAE	R^2
S1+S2	6.65	5.28	0.331	0.352	0.256	0.299
S1+S2+DOP20	**5.64**	**4.38**	**0.519**	**0.274**	**0.201**	**0.574**
DOP20 RGB (10 m)	7.20	5.81	0.217	0.340	0.248	0.343

5.3 Uncertainty Quantification on Tree Height and Density

SegFormer-B3 is evaluated with four uncertainty quantification methods on the test split, using nominal 90% prediction intervals. Point prediction error barely moves with the uncertainty head as height RMSE stays between 6.244 m and 6.298 m, and density RMSE between 0.342 and 0.348 trees per pixel. The comparison is therefore about coverage, interval width, and calibration. On height, the residual-corrected ensemble is the strongest interval model. It reaches coverage of $\mathbf{0.890}$, an ENCE of $\mathbf{0.049}$, and a CRPS of $\mathbf{3.489}$, with an MPIW of 19.95 m. MC dropout with the same residual correction is close in coverage (0.887) but less well calibrated (ENCE 0.102) and slightly wider (MPIW 20.07 m). Split-conformal CQR gives the sharpest height intervals (MPIW $\mathbf{18.92}$ m) with coverage 0.879 and a higher CRPS (3.633). Evidential regression over-covers (0.949) and pays for that with much wider intervals (MPIW 28.07 m) and the worst height calibration (ENCE 0.293). Density intervals are harder to keep at the nominal level. Ensemble and MC dropout both cover 0.867. The ensemble is the sharpest (MPIW $\mathbf{0.871}$) and the most accurate (RMSE $\mathbf{0.342}$), while MC dropout has the highest ENCE (0.300). CQR moves coverage closest to 90% ($\mathbf{0.877}$) with an ENCE of 0.220 and an MPIW of 0.912. Evidential regression reverses its height result as it is the

best calibrated density model (ENCE $\mathbf{0.097}$) and matches the best CRPS ($\mathbf{0.177}$), with coverage 0.926 and the widest intervals (MPIW 1.031).

All methods calibrated on validation data (ensemble, MC dropout, and CQR) fall short of 90% coverage on the test regions for both targets, consistent with a spatial distribution shift between validation and test blocks that violates the exchangeability assumption behind split-conformal guarantees. Patch-level predicted height uncertainty is rank-correlated with the inventory within-cell height spread for all methods (Spearman ρ between 0.47 for the ensemble and 0.68 for evidential regression). See appendix E.5

Table 3: SegFormer-B3 uncertainty on the test split. Intervals are nominal 90%. Ensemble and MC dropout use a validation residual, CQR uses split conformal.

Task	Method	RMSE	Cov.	MPIW	ENCE	CRPS
Height	Ensemble + residual	**6.244**	**0.890**	19.95	**0.049**	**3.489**
	MC dropout + residual	6.298	0.887	20.07	0.102	3.520
	CQR	6.249	0.879	**18.92**	0.065	3.633
	Evidential	6.258	0.949	28.07	0.293	3.600
Density	Ensemble + residual	**0.342**	0.867	**0.871**	0.281	0.178
	MC dropout + residual	0.348	0.867	0.878	0.300	0.180
	CQR	0.344	**0.877**	0.912	0.220	0.180
	Evidential	0.348	0.926	1.031	**0.097**	**0.177**

6 Conclusion

We introduced TreeUQ, a multimodal Earth observation benchmark for estimating tree structure at 10 m resolution across Bavaria. TreeUQ aligns seasonal Sentinel-1 and Sentinel-2 composites, a tree-species raster, and 20 cm DOP20 orthophotos with six per-cell targets derived from an individual-tree inventory, and provides geographically disjoint training, validation, and test regions. To our knowledge, it is the first open benchmark that pairs tree-height statistics with tree-count targets at Sentinel resolution.

For uncertainty quantification, deep ensembles, Monte Carlo dropout, and conformalized quantile regression approach the nominal 90% coverage for height (0.879–0.890), but their intervals remain wide at roughly 19–20 m. Density intervals are harder to calibrate as the same methods reach only 0.867-0.877 coverage, and only evidential regression attains low calibration error, at the cost of over-coverage and the widest intervals. All methods calibrated on validation regions under-cover on held-out test regions, suggesting that spatial distribution shift undermines the exchangeability assumption on which conformal guarantees rely. Predicted height uncertainty is rank-correlated with the inventory within-cell height spread for every method (Spearman $\rho = 0.47$-0.68), so the released variance statistics offer a reference for evaluating uncertainty beyond coverage.

TreeUQ has some limitations, its labels come from one temperate region and density is evaluated only where trees are present. We hope the benchmark supports work on transferring high-resolution information into Sentinel-only models and on uncertainty estimates that remain calibrated under spatial shift.

AI use statement

In this work, we used generative AI tools to refactor Sentinel-1/Sentinel-2 masking and uploading the dataset to huggingface as well as code scaffolding. For interpretation of results, LLMs were used to draft latex of descriptive summaries of the metric tables. The experimental design, the choice of baselines and uncertainty-quantification methods, the evaluation protocol, and all scientific conclusions are entirely the authors' own.

We have not used generative AI tools to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate or refine hypotheses, or to design or provide feedback on our

research methodology or experiments. Additionally, we used generative AI tools for modifying scientific figures and tables, including their formatting

We have reviewed all AI-assisted work. All code that produces reported numbers was read line by line and tested by an author, and the data pipeline was additionally validated by manual inspection of sampled patches against the source rasters and by end-to-end consistency checks on the label statistics reported. All AI-drafted or AI-edited prose was rewritten or approved by the authors, who verified that it makes no claim not supported by our experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects, no personally identifiable information, and no user data, and it therefore required no IRB or equivalent review. All inputs and labels are derived from openly published data: Copernicus Sentinel-1 and Sentinel-2 imagery, the Bavarian DOP20 orthophoto product and a published tree species map. We redistribute these sources under terms that permit redistribution and with attribution.

The intended use of TreeUQ is methodological research on multimodal regression and uncertainty quantification, not operational deployment. We see three risks and address each explicitly. First, geographic scope, the ground truth covers only Bavaria, so models trained on it encode the structure of managed temperate Central European forests and can be miscalibrated elsewhere however generalization methods can be use to train models elsewhere. Second, treating point predictions of height or density as ground truth in carbon accounting, carbon-credit issuance, or regulatory decisions could misinform policy and financial instruments, therefore uncertainty quantification is important. Third, accurate maps of tree height and density could in principle help locate valuable or unmonitored stands for illegal logging. We judge this risk to be low for this release, since the underlying inventory and imagery are already public government open data for the region covered.

REPRODUCIBILITY STATEMENT

All data sources, processing steps, and the release format are described in Section 3 and Appendix A.1–A.2. The spatial split is fixed and distributed with the dataset as index files. The benchmark tasks, masks, and evaluation metrics are defined in Section 4. Architectures, input configurations, and training schedules for all baselines are given in Appendix E, and the DOP20 feature extraction and fusion models in Appendix E. All neural results are reported as mean and standard deviation over three seeds. The dataset is available at the anonymized repository listed in Appendix B, and the code for data processing, training, and evaluation is provided as supplementary material in B

REFERENCES

Steve Ahlswede, Christian Schulz, Christiano Gava, Patrick Helber, Benjamin Bischke, Michael Förster, Florencia Arias, Jörn Hees, Begüm Demir, and Birgit Kleinschmit. *TreeSatAI Benchmark Archive* : a multi-sensor, multi-label dataset for tree species classification in remote sensing. *Earth System Science Data*, 15(2):681–695, February 2023. ISSN 1866-3516. doi: 10.5194/essd-15-681-2023. URL https://essd.copernicus.org/articles/15/681/2023/.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep Evidential Regression. 2020.

Laura Bader, Muhammad Ammar Ahmed, Xiao Xiang Zhu, and Göran Kauermann. Deep Evidential Regression for Sparse Forest Height Estimation from Multimodal Satellite Imagery, August 2026. URL https://arxiv.org/abs/2608.06406v2.

Lukas Blickensdörfer, Katja Oehmichen, Dirk Pflugmacher, Birgit Kleinschmit, and Patrick Hostert. National tree species mapping using Sentinel-1/2 time series and German National Forest Inventory data. *Remote Sensing of Environment*, 304:114069, April 2024. ISSN 0034-4257. doi: 10.1016/j.rse.2024.114069. URL https://www.sciencedirect.com/science/article/pii/S0034425724000804.

John Brandt and Fred Stolle. A global method to identify trees outside of closed-canopy forests with medium-resolution satellite imagery. *International Journal of Remote Sensing*, 42(5):1713–1737, March 2021. ISSN 0143-1161, 1366-5901. doi: 10.1080/01431161.2020.1841324.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL https://dl.acm.org/doi/10.1145/2939672.2939785.

European Space Agency. Sentinel-1 mission overview and data products, 2026a. URL https://documentation.dataspace.copernicus.eu/Data/SentinelMissions/Sentinel1.html.

European Space Agency. *Sentinel-2 Mission Documentation*. Copernicus Data Space Ecosystem, 2026b. URL https://documentation.dataspace.copernicus.eu/Data/SentinelMissions/Sentinel2.html. Accessed: 2026-05-07.

Maximilian Freudenberg, Paul Magdon, and Nils Nölke. Individual tree crown delineation in high-resolution remote sensing images based on U-Net. *Neural Computing and Applications*, 34(24): 22197–22207, December 2022. ISSN 0941-0643, 1433-3058. doi: 10.1007/s00521-022-07640-4. URL https://link.springer.com/10.1007/s00521-022-07640-4.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Yuchang Jiang and Maxim Neumann. Not every tree is a forest: Benchmarking forest types from satellite remote sensing. In *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*, pp. 808–813, 2025. doi: 10.1109/IGARSS55030.2025.11243780.

Patrick Kacic, Frank Thonfeld, Ursula Gessner, and Claudia Kuenzer. Forest Structure Characterization in Germany: Novel Products and Analysis Based on GEDI, Sentinel-1 and Sentinel-2 Data. *Remote Sensing*, 15(8):1969, January 2023. ISSN 2072-4292. doi: 10.3390/rs15081969. URL https://www.mdpi.com/2072-4292/15/8/1969.

Heini Kujala, Francesco Minunno, Virpi Junttila, Ninni Mikkonen, Annikki Mäkelä, Raimo Virkkala, Anu Akujärvi, Niko Leikola, and Risto K. Heikkinen. Role of data uncertainty when identifying important areas for biodiversity and carbon in boreal forests. *Ambio*, 52(11):1804–1818, November 2023. ISSN 0044-7447. doi: 10.1007/s13280-023-01908-2. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10562324/.

Landesamt für Digitalisierung. Digitales orthophoto 20 cm bodenauflösung rgb (dop20 rgb), 2025a. URL https://geodaten.bayern.de/opengeodata/OpenDataDetail.html?pn=dop20rgb.

Landesamt für Digitalisierung. Einzelbäume [Dataset]. Bayerische Vermessungsverwaltung, 2025b. URL https://geodaten.bayern.de/opengeodata/OpenDataDetail.html?pn=einzelbaeume. Accessed: 2026-05-07. License: CC BY 4.0.

Nico Lang, Nikolai Kalischek, John Armston, Konrad Schindler, Ralph Dubayah, and Jan Dirk Wegner. Global canopy height regression and uncertainty estimation from GEDI LiDAR waveforms with deep ensembles. *Remote Sensing of Environment*, 268:112760, 2022. doi: 10.1016/j.rse.2021.112760. URL https://doi.org/10.1016/j.rse.2021.112760.

Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the Earth. *Nature Ecology & Evolution*, 7(11):1778–1789, November 2023. ISSN 2397-334X. doi: 10.1038/s41559-023-02206-6. URL https://www.nature.com/articles/s41559-023-02206-6.

Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540, 2022.

Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112:102926, August 2022a. ISSN 1569-8432. doi: 10.1016/j.jag.2022.102926. URL https://www.sciencedirect.com/science/article/pii/S1569843222001248.

Jingbo Li, Changchun Li, Weimeng Xu, Haikuan Feng, Fa Zhao, Huiling Long, Yang Meng, Weinan Chen, Hao Yang, and Guijun Yang. Fusion of optical and SAR images based on deep learning to reconstruct vegetation NDVI time series in cloud-prone regions. *International Journal of Applied Earth Observation and Geoinformation*, 112:102818, August 2022b. ISSN 1569-8432. doi: 10.1016/j.jag.2022.102818. URL https://www.sciencedirect.com/science/article/pii/S1569843222000206.

Meixuan Li, Guoqing Wang, Tianyu Li, Yang Yang, Wei Li, Xun Liu, and Ying Liu. Knowledge-Guided Multi-Task Network for Remote Sensing Imagery. *Remote Sensing*, 17(3):496, January 2025. ISSN 2072-4292. doi: 10.3390/rs17030496. URL https://www.mdpi.com/2072-4292/17/3/496.

Min Lu, Jiayin Liu, Feng Wang, and Yuming Xiang. Multi-Task Learning of Relative Height Estimation and Semantic Segmentation from Single Airborne RGB Images. *Remote Sensing*, 14 (14):3450, January 2022. ISSN 2072-4292. doi: 10.3390/rs14143450. URL https://www.mdpi.com/2072-4292/14/14/3450.

David Morin, Milena Planells, Stéphane Mermoz, and Florian Mouret. Estimation of forest height and biomass from open-access multi-sensor satellite imagery and GEDI Lidar data: high-resolution maps of metropolitan France, January 2024. URL https://zenodo.org/doi/10.5281/zenodo.8071003.

Yang Mu, Jianhua Guo, Muhammad Shahzad, and Xiao Xiang Zhu. National-scale tree species mapping with deep learning reveals forest management insights in Germany. *International Journal of Applied Earth Observation and Geoinformation*, 139:104522, 2025. doi: 10.1016/j.jag.2025.104522.

Peter Potapov, Xinyuan Li, Andres Hernandez-Serna, Alexandra Tyukavina, Matthew C. Hansen, Anil Kommareddy, Amy Pickens, Svetlana Turubanova, Hao Tang, Carlos Edibaldo Silva, John Armston, Ralph Dubayah, J. Bryan Blair, and Michelle Hofton. Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sensing of Environment*, 253: 112165, February 2021. ISSN 0034-4257. doi: 10.1016/j.rse.2020.112165. URL https://www.sciencedirect.com/science/article/pii/S0034425720305381.

Yujun Quan, Rongrong Zhang, Jian Li, Song Ji, Hengliang Guo, and Anzhu Yu. Learning SAR-Optical Cross Modal Features for Land Cover Classification. *Remote Sensing*, 16(2):431, January 2024. ISSN 2072-4292. doi: 10.3390/rs16020431. URL https://www.mdpi.com/2072-4292/16/2/431.

Rahul Rahaman and Alexandre H. Thiery. Uncertainty quantification and deep ensembles. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

Raphael Rossellini, Rina Foygel Barber, and Rebecca Willett. Integrating uncertainty awareness into conformalized quantile regression. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1540–1548. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/rossellini24a.html.

Swalpa Kumar Roy, Ankur Deria, Danfeng Hong, Behnood Rasti, Antonio Plaza, and Jocelyn Chanussot. Multimodal Fusion Transformer for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2023. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2023.3286826.

Martin Schwartz, Philippe Ciais, Catherine Ottlé, Aurelien De Truchis, Cedric Vega, Ibrahim Fayad, Martin Brandt, Rasmus Fensholt, Nicolas Baghdadi, François Morneau, David Morin, Dominique Guyon, Sylvia Dayau, and Jean-Pierre Wigneron. High-resolution canopy height map in the Landes forest (France) based on GEDI, Sentinel-1, and Sentinel-2 data with a deep learning approach. *International Journal of Applied Earth Observation and Geoinformation*, 128:103711, April 2024. ISSN 15698432. doi: 10.1016/j.jag.2024.103711.

Yang Su, Martin Schwartz, Ibrahim Fayad, Mariano García, Miguel A. Zavala, Julián Tijerín-Triviño, Julen Astigarraga, Verónica Cruz-Alonso, Siyu Liu, Xianglin Zhang, Songchao Chen, François Ritter, Nikola Besic, Alexandre d'Aspremont, and Philippe Ciais. Canopy height and biomass distribution across the forests of Iberian Peninsula. *Scientific Data*, 12(1):678, April 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05021-9. URL https://www.nature.com/articles/s41597-025-05021-9.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Carlos Gomes, Ankur Kumar, Vishal Gaur, Myscon Truong, Denys Godwin, Sam Khallaghi, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Disha Shidham, Rufai Omowunmi Balogun, Venkatesh Kolluru, Trevor Keenan, Paulo Arevalo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Timothy Mayer, Christopher Hain, Robert Kennedy, Bianca Zadrozny, David Bell, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-eo-2.0: A versatile multitemporal foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026. doi: 10.1109/TGRS.2025.3642610.

Fei Tong and Yun Zhang. Individual tree crown delineation in high resolution aerial RGB imagery using StarDist-based model. *Remote Sensing of Environment*, 319:114618, March 2025. ISSN 0034-4257. doi: 10.1016/j.rse.2025.114618. URL https://www.sciencedirect.com/science/article/pii/S0034425725000227.

Jonas Troles, Ute Schmid, Wen Fan, and Jiaojiao Tian. BAMFORESTS: Bamberg Benchmark Forest Dataset of Individual Tree Crowns in Very-High-Resolution UAV Images. *Remote Sensing*, 16 (11):1935, May 2024. ISSN 2072-4292. doi: 10.3390/rs16111935. URL https://www.mdpi.com/2072-4292/16/11/1935.

Bayerische Vermessungsverwaltung. Opendata einzelbäume, 2025. https://geodaten.bayern.de/opengeodata/OpenDataDetail.html?pn=einzelbaeume [Accessed: 2026-04-14].

Nan Wang, Donghui Xie, Lin Jin, Yi Li, Xihan Mu, and Guangjian Yan. Unveiling Forest Density Dynamics in Saihanba Forest Farm by Integrating Airborne LiDAR and Landsat Satellites. *Remote Sensing*, 17(19):3338, January 2025. ISSN 2072-4292. doi: 10.3390/rs17193338. URL https://www.mdpi.com/2072-4292/17/19/3338.

Mareike Weishaupt, Wen Fan, Jonas Troles, and Jiaojiao Tian. Individual Tree Crown Based Tree Species Classification From Very High-Resolution UAV Images. 2025.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: simple and efficient design for semantic segmentation with transformers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation, October 2025. URL http://arxiv.org/abs/2403.15356. arXiv:2403.15356 [cs.CV].

Ling Yao, Tang Liu, Jun Qin, Ning Lu, and Chenghu Zhou. Tree counting with high spatial-resolution satellite imagery based on deep neural networks. *Ecological Indicators*, 125:107591, June 2021. ISSN 1470-160X. doi: 10.1016/j.ecolind.2021.107591. URL https://www.sciencedirect.com/science/article/pii/S1470160X21002569.

A Documentation and Uses

A.1 Data Sources and Processing

Temporal coverage. All Sentinel composites cover a single year. Seasons are defined as spring (2025-03-01 to 2025-05-31), summer (2025-06-01 to 2025-08-31), autumn (2025-09-01 to 2025-11-30), and winter (2025-12-01 to 2026-02-28). DOP20 chips stem from the two-year Bavarian flight cycle, in which northern and southern Bavaria are flown in alternate years. The 1 km source orthophotos underlying the released chips were flown between May 2023 and September 2025, all during the leaf-on months May to September.

Sentinel-2. Level-2A scenes were retrieved from the Microsoft Planetary Computer (`sentinel-2-l2a`), keeping scenes with scene-level cloud cover below 20% and at most 500 scenes per season. Before compositing, pixels with Scene Classification Layer classes 1 (saturated or defective), 3 (cloud shadow), 8 and 9 (cloud, medium and high probability), 10 (thin cirrus), and 11 (snow or ice) are masked. The composite is the per-pixel median over the remaining observations. Bands are stored in the order B2, B3, B4, B8, B5, B6, B7, B8A, B11, B12 while the native 20 m bands are bilinearly resampled to 10 m in EPSG:25832. Values are L2A digital numbers.

Sentinel-1. Up to 200 GRD scenes per season are retrieved from `sentinel-1-grd` and restricted to ascending passes. VV and VH are retained and non-positive and no-data pixels are masked, and the composite is the per-pixel median, bilinearly reprojected to 10 m.

Missing values. No temporal gap-filling or imputation is performed. Sentinel-2 pixels masked on every date of a season remain NaN. Sentinel-1 is nearly complete i.e fewer than 0.01% of pixels, confined to 369 patches at the margin of the acquisition footprint, are NaN, identically in all four seasons. Height targets (mean, median, and variance) are NaN exactly in cells without inventory trees

Tree species. The species layer is a 10 m raster (EPSG:3035), reprojected to EPSG:25832 with nearest-neighbour resampling and snapped to the reference grid. It is stored as raw `uint8` class IDs in a single channel. Class shares over the released patches are given in Table 4.

Table 4: Species classes and their share of all pixels in the 45,419 released patches.

ID	Class	Share of pixels
0	Beech	1.77%
1	Douglas fir	0.61%
2	Fir	0.49%
3	Larch	0.46%
4	Oak	0.42%
5	Other deciduous	1.68%
6	Pine	6.55%
7	Spruce	8.77%
11	No tree / no-data	79.27%

DOP20 coverage. A 6400×6400 chip is released when at least 50% of its footprint contains data. Of the 45,419 patches, 43,822 (96.5%) have a chip, 42,872 of which are essentially complete ($\geq$99.9% coverage).

Size. The core component (Sentinel-1, Sentinel-2, species, and labels) occupies 150.8 GiB across 837 tiles and the DOP20 component occupies 5,118.1 GiB, for a total of 5.15 TiB.

A.2 ORGANIZATION

The release is partitioned into 837 geographic tiles, each corresponding to a 10×10 km cell in EPSG:25832 and named after the kilometre coordinates of its lower-left corner. A tile holds up to 64 patches. Each tile stores the core tensors and the DOP20 chips in separate archives, so the core component can be downloaded without the aerial imagery. Within an archive, the members of a patch share a zero-padded `patch_id` prefix, and `schema.json` lists the dtype, shape, and band names of every member. The split parquet files map each patch to its tile and archive paths and carry its split, block, and DOP20 availability.

```
<OUTPUT_DIR>/
├── README.md
├── croissant.json
└── data/
    ├── schema.json
    ├── index/
    │   ├── train.parquet
    │   ├── validation.parquet
    │   └── test.parquet
    └── tiles/
        ├── tile_490_5520/
        │   ├── core.tar
        │   ├── dop20.tar
        │   └── manifest.json
        ├── tile_490_5530/
        └── ...
```

A.3 INTENDED USE

The TreeUQ dataset is intended as a research and benchmarking dataset for developing and evaluating machine learning models in Earth observation. In particular, it supports methods for multimodal fusion (Sentinel-1 and Sentinel-2), multi-task learning (joint prediction of tree count and height), and uncertainty-aware modeling. The dataset is designed for methodological research, not for direct operational deployment without additional validation.

A.4 USE CASES

The dataset enables several core use cases:

- **Multi-task learning**: Joint as well as individual prediction of heterogeneous targets from shared inputs.
- **Forest structure estimation**: Large-scale inference of tree density and height from medium-resolution satellite imagery.
- **Cross-modal knowledge distillation**: Distill knowledge using 20cm high resolution rgb with 10m sentinel data.
- **Uncertainty modeling**: Evaluation of models that predict both mean values and associated variability.
- **Spatial generalization**: Studying robustness and out-of-distribution performance using spatially disjoint splits.
- **Domain Adaptation and Zero-Shot Transfer**: Because the Sentinel-1 and Sentinel-2 input modalities are globally consistent, models trained on this dataset can be tested for out-of-distribution (OOD) generalization and unsupervised domain adaptation in different global biomes.
- **Benchmarking EO models**: Standardized comparison of classical models, deep learning architectures, and foundation models.

A.5 LIMITATIONS

The dataset has several limitations:

- **Label sparsity**: Ground truth is sparse and unevenly distributed, which may bias learning.
- **Resolution constraints**: At 10 m resolution, targets must be inferred indirectly as individual trees are not observable. 20cm high resolution RGB data can be used for knowledge distillation.
- **Geographic scope**: The ground-truth dataset is limited to the federal state of Bavaria. The models may suffer from domain shift if applied directly to vastly different biomes (e.g., tropical rainforests). However, because the satellite inputs are global, this presents an excellent open challenge for the community regarding domain adaptation.
- **Temporal aggregation**: Seasonal composites remove short-term dynamics and may obscure temporal variability.

A.6 DOCUMENTATION AND MAINTENANCE

- **Versioning**: Detailed version history of the dataset will be maintained to track changes and improvements over time.
- **Version Control**: The benchmark will be versioned to ensure transparency in updates, including bug fixes and task expansions.
- **Community Feedback**: We welcome contributions and suggestions from the community to expand, refine, and validate the benchmark, fostering collaborative progress in building robust AI systems for Earth observation.
- **Long-term Preservation**: The dataset is archived in Huggingface to ensure long-term availability.

A.7 BROADER IMPACTS

- **Positive Impacts**: The automated, large-scale estimation of tree density and height from freely available satellite imagery has significant positive societal implications. It provides ecologists, climate scientists, and policymakers with scalable tools to accurately monitor above-ground biomass, track deforestation or afforestation efforts, and improve global carbon accounting. By democratizing access to high-resolution forest metrics without the need for expensive airborne LiDAR campaigns, this work supports global conservation initiatives and sustainable forest management.
- **Negative Impacts and Risks of Misuse**: If predictive models trained on this dataset are deployed blindly in geographically distinct regions (e.g., tropical or boreal forests) without proper domain adaptation, it could lead to highly inaccurate carbon accounting. This could misinform conservation policies or unfairly skew carbon credit markets. Accurate maps of tree density and height could theoretically be misused by malicious actors to identify lucrative, unmonitored forest regions for illegal logging. Finally, in regulatory or financial contexts, treating automated point predictions as absolute truth, without factoring in the inherent uncertainties of remote sensing, can lead to harmful real-world decisions. We actively attempt to mitigate this final risk by providing per-pixel variance statistics within the dataset, explicitly encouraging the machine learning community to develop uncertainty-aware predictive models.

B ACCESSIBILITY

The TreeUQ dataset can be accessed and downloaded through our Huggingface repository (https://huggingface.co/datasets/iclr2027kiwi/TreeUQ). The code repository for data processing and model training is present here: https://anonymous.4open.science/r/treeuq-DF65/README.md

C Author Statement and Data License

- Author Responsibility Statement: The authors bear all responsibilities in case of any violations of rights or ethical concerns regarding the TreeUQ dataset.
- Data License Confirmation: The dataset is released under the CC-BY-4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

D Hosting and Accessibility

To ensure persistent accessibility and long-term preservation, the TreeUQ dataset is publicly hosted on the Hugging Face. The authors are committed to the continuous curation of the dataset, with updates scheduled bi-annually or as required by significant modifications to the primary data sources.

E Model Implementation

E.1 Compute Resources

All experiments were conducted on a single compute node equipped with 5 Tesla P100, 2 V100 and 1 V100s GPU and 48 CPU cores and 380 GB of RAM.

E.2 Modality Ablations

We also conduct a controlled ablation study to quantify the contribution of individual modalities. We progressively evaluate:

- All four Sentinel-2 seasons
- All four Sentinel-1 seasons
- Combined Sentinel-1 and Sentinal-2
- Combined Sentinel-1, Sentinel-2 and Tree Species

A height estimation modality ablation shows that Sentinel-2 has a lot of the usable height signal. Training on S2 only achieved the best error, with an RMSE of $\mathbf{6.118}$ m and an MAE of $\mathbf{4.776}$ m. Adding Sentinel-1 left performance essentially unchanged (RMSE 6.133 m, MAE 4.799 m), confirming that C-band SAR contributes little once optical features are available. Further concatenating the species map slightly increased error on this seed (RMSE 6.233 m, MAE 4.886 m), so the extra class channel does not improve the point estimate under this protocol. In contrast, Sentinel-1 alone is a weak predictor (RMSE 7.475 m, MAE 6.143 m), indicating that radar backscatter without optical context is insufficient for inventory mean height at 10 m. Optical spatial structure is therefore the critical modality.

Table 5: Height-only UNet modality ablation on the original scale.

Inputs	RMSE	MAE
S1+S2+species	6.233	4.886
S1+S2	6.133	4.799
S2 only	**6.118**	**4.776**
S1 only	7.475	6.143

E.3 DOP20 Experiment Details

Feature cache. Each $6{,}400 \times 6{,}400$ orthophoto is divided into a 128×128 grid of non-overlapping 50×50-pixel tiles, one per 10 m Sentinel cell. A frozen ImageNet-pretrained ResNet50 encodes each tile into a 2048-dimensional vector (global average pooling of the final stage). Incremental PCA, fitted on training patches only, reduces these vectors to 256 components. The resulting maps

are stored as `float16` tensors of shape $256 \times 128 \times 128$ for the training, validation, and test splits. The cache is computed offline and is not updated during training.

S1+S2+DOP20 model. Sentinel-1 and Sentinel-2 inputs are processed by separate ImageNet-pretrained ResNet50 encoders. The cached DOP20 maps are not re-encoded, rather they are pooled to the resolution of each U-Net skip connection and fused with the Sentinel features through 1×1 convolutions.

DOP20 RGB model. Each 50×50-pixel block of the orthophoto is averaged, producing a $3 \times 128 \times 128$ input at 10 m. A standard U-Net with an ImageNet-pretrained ResNet50 encoder is trained on this input alone, without Sentinel data.

Test-time inputs. All settings use DOP20 at test time: the test loader concatenates the same 256-channel cache with S1+S2 for the fused model, and reads the 10 m RGB cache for the RGB-only model.

E.4 Evaluated Architectures and Hyperparameters

We evaluated five baselines under a fixed computational budget of 60 effective epochs:

- **XGBoost**: Serving as a pixel-wise, non-spatial baseline, we trained two independent XGBoost models (one per target) using 2,000 estimators, a maximum depth of 6, and a learning rate of 0.03. To maintain tractability, the model was trained on a 3% random subsample of valid training pixels.
- **U-Net (ResNet-50)**: A standard U-Net with an ImageNet-pretrained ResNet-50 encoder. The multi-seasonal Sentinel stack (48 channels) was ingested by modifying the initial convolutional layer. The network was optimized using AdamW and a global batch size of 64. A Cosine Annealing learning rate scheduler was applied.
- **SegFormer-B3**: Utilizing a PVTv2-B3 encoder and an all-MLP SegFormer-style decoder. It was trained using AdamW, and a batch size of 8 on a single GPU.
- **Prithvi 2.0**: The Prithvi-EO-2.0-300M ViT-L encoder (24 layers, embedding dimension 1024, spatio-temporal patch size $1 \times 16 \times 16$), initialized from the official pretrained weights. Because Prithvi was pretrained on HLS optical data, Sentinel-1 was discarded and the six HLS-equivalent Sentinel-2 bands (B02, B03, B04, B8A, B11, B12) from each of the four seasons were fed as a 4-frame time series, normalized with Prithvi's pretraining statistics. Training had two phases: first a frozen encoder (30 epochs, learning rate 10^{-4}, global batch size 48), then full fine-tuning starting from the best Phase 1 checkpoint (30 epochs, learning rate 5×10^{-5}, global batch size 48). Both phases used AdamW (weight decay 10^{-4}), cosine annealing, mixed precision, and DDP on 3 GPUs.
- **DOFA**: The DOFA ViT-Large encoder, initialized from the official pretrained weights, with positional embeddings interpolated to the 128×128 input. DOFA's wavelength-conditioned dynamic patch embedding allows Sentinel-1 and Sentinel-2 to be ingested together. Each season was encoded separately as a 12-channel image (VV, VH, and the 10 Sentinel-2 bands), with the central wavelengths passed to the encoder. The two-phase schedule was also the same: a frozen encoder (30 epochs, learning rate 10^{-4}, global batch size 48), then full fine-tuning (30 epochs, learning rate 5×10^{-5}, global batch size 48). Training used AdamW, cosine annealing, mixed precision, gradient checkpointing, and DDP on 3 GPUs.

E.5 Predicted uncertainty and inventory height variance

Figure below compares each method's patch-mean predicted uncertainty with the patch-mean inventory scale $\sqrt{\text{height_variance}}$. The ρ is the Spearman rank correlation between the two patch means. The inventory variance is a neighbourhood statistic of the tree inventory, and a high ρ means the model assigns larger uncertainty where that statistic is larger.

The deep ensemble ($\rho = 0.47$) and MC dropout ($\rho = 0.57$) form flat, dense bands at small predicted σ, mostly below about 3 m, with little slope against the inventory scale.

Conformalized quantile regression ($\rho = 0.67$) shows that the patch-mean σ increases from roughly $2\,\mathrm{m}$ to about $8\,\mathrm{m}$ as $\sqrt{\text{height_variance}}$ increases.

The evidential head has a similar rank correlation ($\rho = 0.68$) but a different magnitude. Its patch-mean σ extends to about 15–$17\,\mathrm{m}$ while the inventory scale remains below about $2.2\,\mathrm{m}$.

