ECHOSAT: ESTIMATING CANOPY HEIGHT OVER SPACE AND TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Forest monitoring is critical for climate change mitigation. However, existing global tree height maps only provide static snapshots and do not capture temporal forest dynamics, which are essential for accurate carbon accounting. We introduce ECHOSAT, a global and temporally consistent tree height map at $10\,\mathrm{m}$ resolution spanning multiple years, which accurately quantifies tree growth and disturbances over time. To this end, we resort to multi-sensor satellite data to train a vision transformer model, which performs pixel-level temporal regression. A self-supervised growth loss regularizes the predictions to follow growth curves that are in line with natural tree development, including gradual height increases over time, but also abrupt declines due to forest loss events such as fires. Our evaluation further shows that our model improves state-of-the-art accuracies in the context of single-year predictions. ECHOSAT can be accessed here: `https://echosatanonymous.projects.earthengine.app/view/echosat-temporal`.

1 INTRODUCTION

Forests play a crucial role in the mitigation of climate change, absorbing about $3.5\,\mathrm{Pg}$ of carbon per year, which represents almost half of the anthropogenic fossil fuel emissions worldwide (Pan et al., 2024). As global carbon emissions continue to increase, precise monitoring of forest carbon dynamics using up-to-date information on forest health and carbon balance has become essential for effective climate policy and forest management decisions. Recent advances in satellite remote sensing and machine learning have enabled automated forest carbon monitoring on country-to-global scales, using tree height as a key proxy for estimating the so-called *above-ground biomass* (AGB) and, hence, carbon storage (Schwartz et al., 2023).

Most of these height maps, however, provide a static representation of forests at a specific point in time and cannot be used to estimate year-to-year carbon absorption (Lang et al., 2023; Pauls et al., 2024; Potapov et al., 2021; Tolan et al., 2024). While such static global snapshots of forests are a valuable resource, they do not capture temporal dynamics such as tree growth or forest loss. Some studies monitor changes of forests over time, but they are often limited to large-scale disturbances such as forest loss caused by major wildfires (Hansen et al., 2013b; Reiche et al., 2021). Small-scale height decreases from degradation, individual tree mortality or forest thinning, however, are significantly smaller and, hence, harder to detect. Additionally, very few studies have succeeded in retrieving realistic year-to-year forest growth pattern at a high resolution (Schwartz et al., 2025; Turubanova et al., 2023), and rely on single-year models independently applied to multiple years along with extensive post-processing to achieve temporal consistency. None of the these approaches is based on models that inherently learn forest temporal dynamics, potentially leading to unrealistic fluctuations at pixel level and poor temporal coherence in predictions when no post-processing is applied.

We provide the first global tree height mapping approach at high resolution ($10\,\mathrm{m}$) across multiple years. Our method trains a state-of-the-art transformer-based temporal regression model (modified to fit our data modalities) with an adapted loss that addresses sparse supervision in both space (not every pixel is associated with a ground-truth label) and time (most pixels are associated with only few measurements over the years). More precisely, our contributions are threefold: First, we present *ECHOSAT*, the first high-resolution ($10\,\mathrm{m}$) spatio-temporal tree height map covering the entire globe across seven years (2018–2024), which paves the way for reliable monitoring of forest dynamics and disturbances at scale. Second, to enforce physically plausible forest growth patterns, we introduce a

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.

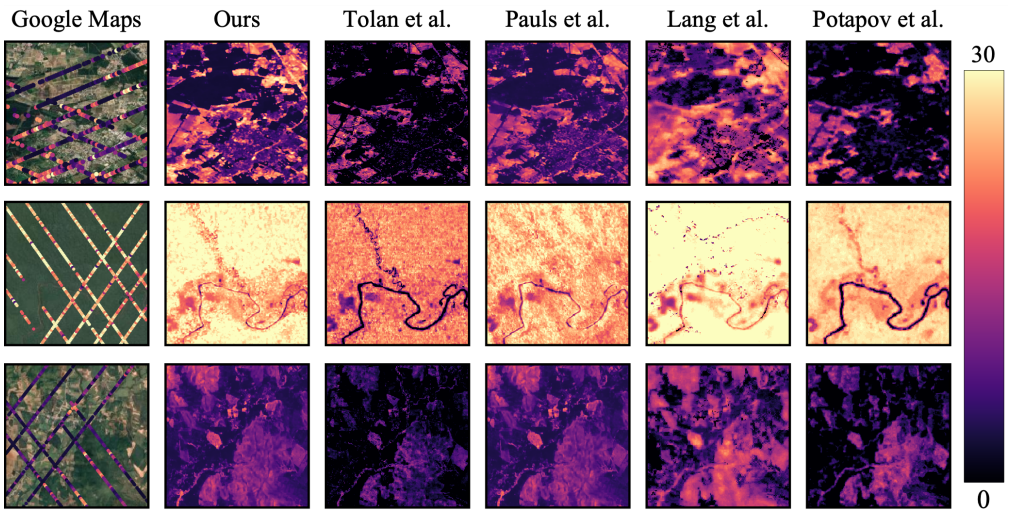


Figure 1: Qualitative comparison across three locations. The first column shows Google Maps imagery along with (sparse) GEDI height measurements, while the other columns show predicted tree heights ($0\,\mathrm{m}-30\,\mathrm{m}$ range) for various methods. The comparison highlights differences in spatial detail, forest boundary detection, and height estimation accuracy between the approaches.

novel growth loss framework that is specifically designed for training temporal regression models with sparse ground truth labels, both in space and time. Third, we demonstrate that our model inherently learns realistic temporal forest height dynamics without relying on post-processing, capturing both natural growth and abrupt disturbances.[1]

2 Background

Remote sensing has long been used to complement and upscale forest inventory measurements (Tomppo et al., 2008), and more recently deep learning approaches have been introduced in this context. We briefly review these methods and highlight the relevance and impact of our work.

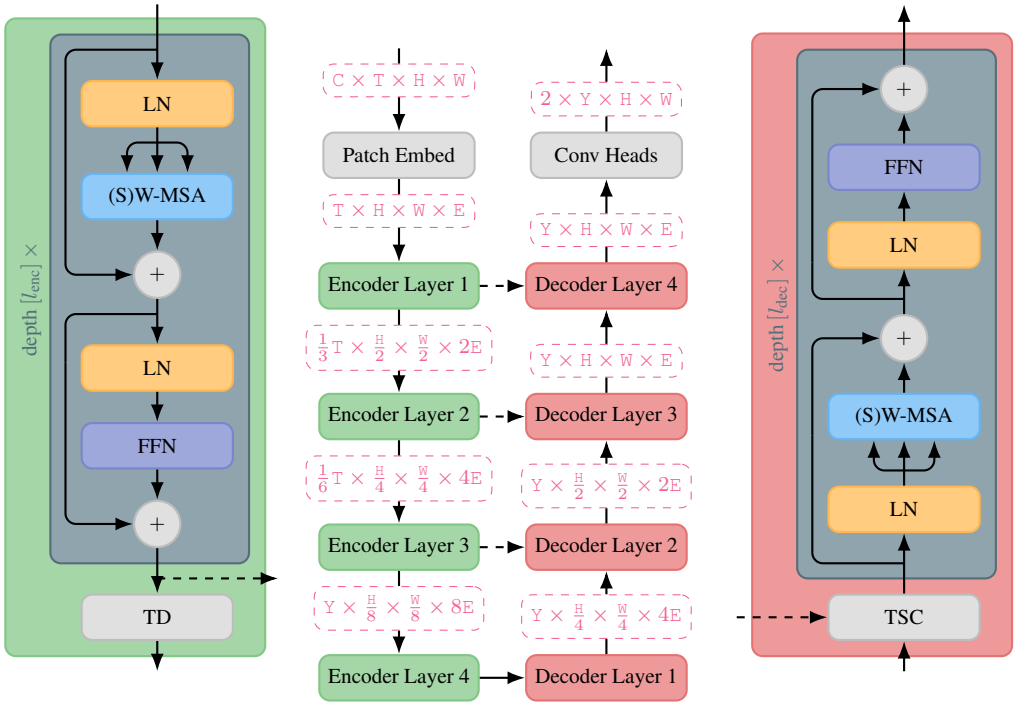
2.1 Remote Sensing Data

High-resolution satellite remote sensing mainly employs three sensor types: optical, SAR (Synthetic Aperture Radar) and LiDAR (Light Detection And Ranging). Optical sensors operate passively, measuring the sun's reflected electromagnetic radiation across multiple spectral bands from visible to near-infrared wavelengths. For instance, Sentinel-2 delivers multi-spectral optical imagery with up to $10\,\mathrm{m}$ spatial resolution and an approximately five-day revisit time depending on latitude, while Landsat provides historical multi-spectral data with $30\,\mathrm{m}$ spatial resolution, enabling long-term analyses. In contrast, SAR sensors actively transmit microwave signals and measure the backscattered energy, enabling data acquisition regardless of illumination and cloud cover. LiDAR sensors are light-emitting and receiving sensors that estimate distances by measuring the time it takes for the light to return to the sensor after being reflected on an object. The Global Ecosystem Dynamics Investigation (GEDI) mission, operated by the NASA and deployed on the International Space Station (ISS), provides spaceborne LiDAR measurements of forest vertical structure within $25\,\mathrm{m}$ diameter footprints since the end of 2018 (Dubayah et al., 2022). These measurements can provide (above-ground) height estimates of the footprint (see Figure 1 left column). GPS and star tracker data are used to estimate the position of the ISS and to deduce the geolocation of a measurement.

2.2 Vision (Foundation) Architectures

Recent advances in remote sensing have been driven by foundation models adapted from computer vision (Astruc et al., 2025; Fayad et al., 2025; Fuller et al., 2023; Tseng et al., 2023; 2025). These

[1] At the same time, our model outperforms existing approaches in single-year evaluations.



(a) Encoder Layer l_{enc} (b) Temporal-Swin-Unet architecture (c) Decoder Layer l_{dec}

Figure 2: Illustrations of the Temporal-Swin-Unet architecture. Skip connections are depicted as dashed arrows. The shape of the tensor in between a layer is shown in magenta: C (channels), T (timesteps), H (height), W (width), Y (years), and E (embedding dimension). In addition to the Video Swin Transformer Blocks (Liu et al., 2022), Encoder Layers have a *Temporal Downsample* (TD) layer at the end, and Decoder Layers a *Temporal Skip Connection* (TSC) at the beginning.

models predominantly employ isotropic architectures (e.g., standard vision transformers) to leverage scalable self-supervised learning (SSL) objectives. The three dominant paradigms are: (1) Masked Image Modeling (MIM), such as MAE (He et al., 2022), which learns by reconstructing randomly masked patches; (2) Contrastive Learning, such as SimCLR (Chen et al., 2020), which optimizes for semantic invariance between different views or crops of an image; and (3) Joint-Embedding Prediction Architectures like I-JEPA (Assran et al., 2023) that predict the embedding of one part of the image using another part of the same image.

However, these paradigms present structural incompatibilities with hierarchical architectures (Cao et al., 2022; Liu et al., 2021), which are otherwise superior for dense prediction tasks. Specifically, the unstructured masking strategies central to MAE and I-JEPA disrupt the rigid grid alignment required by the shifted-window attention mechanisms in hierarchical transformers. Furthermore, contrastive objectives often enforce global semantic uniformity, suppressing the high-frequency, pixel-level spatial details required for fine-grained regression. Consequently, while Swin-based architectures offer inductive biases well-suited for dense canopy height estimation, they have seen limited adoption in large-scale remote sensing foundation models.

2.3 Forest Height Prediction

Due to the correlation between forest height and forest biomass, forest height mapping has gained significant attention in recent years, with numerous studies producing tree height maps at regional (Favrichon et al., 2025), national (Schwartz et al., 2023; Su et al., 2025), continental (Liu et al., 2023) and global (Lang et al., 2023; Pauls et al., 2024; Potapov et al., 2021; Tolan et al., 2024) scales. These maps typically combine remote sensing imagery with reference height measurements from spaceborne LiDAR systems such as GEDI or ICESat, or from airborne laser scanning (ALS) campaigns. The development of global tree height maps has progressed significantly in recent years. Potapov et al. (2021) pioneered the first global tree height map using Landsat data at 30 m resolution, GEDI

measurements, and a random forest model. Subsequent work by Lang et al. (2023) improved the spatial resolution to 10 m using Sentinel-2 data and convolutional neural networks. More recently, Pauls et al. (2024) developed a global map using a U-Net architecture with a specialized loss function designed to improve robustness to noise, while Tolan et al. (2024) achieved individual tree-level detection using a DINOv2 model fine-tuned on 1 m+ Maxar data with ALS and GEDI labels. Figure 1 shows three spots with these canopy height maps and corresponding GEDI measurements.

Single-snapshot estimates cannot capture the effects of management or climate change over time. Temporal tree height mapping methods address this, with the simplest approach training single-year models independently on each year of remote sensing data (Kacic et al., 2023). A more sophisticated approach employs space-for-time substitution, where models trained on spatial variations are applied to temporal sequences under the assumption that similar spatial patterns correspond to similar temporal dynamics (Schwartz et al., 2025). A third approach uses classical machine learning methods with extensive post-processing to smooth temporal inconsistencies and reduce prediction uncertainty (Turubanova et al., 2023). These approaches easily capture abrupt large-scale height changes due to forest clearcuts or large disturbance events (e.g., fires, storms) but often overlook small disturbances at tree level. Above all, they struggle to generate accurate and temporally consistent pixel-level height time series, limiting their applicability for detailed carbon dynamics analyses.

3 Approach

We propose an approach to estimate tree height with coherent temporal predictions at a global scale by training models to capture realistic temporal dynamics. Our approach comprises a model with two output heads, trained incrementally, see Figure 3: First, we train a *reference head* using the Huber loss. In essence, this head yields height predictions for individual years that may be physically implausible (e.g., unrealistic fluctuations in tree height due data noise). Next, a *prediction head* is trained using pseudo-labels from the reference head and physically motivated constrained (piecewise) linear regression models. We outline the approach below and refer to Appendix A.1 for details.

3.1 Data

We integrate multi-temporal satellite data spanning 2018–2024 to enable global-scale temporal tree height mapping, with Sentinel-2 providing the primary image source at 10 m resolution.

Multi-sensor Satellite Data. We combine optical (Sentinel-2) and radar backscatter (Sentinel-1, ALOS PALSAR-2) data with auxiliary products (TanDEM-X DEM and forest classification). Sentinel-2 provides monthly images across 12 spectral bands (10 m−60 m native resolution, resampled to 10 m), while radar data offer quarterly (Sentinel-1) and yearly (ALOS PALSAR-2) composites. GEDI LiDAR measurements serve as ground-truth labels for 2019–2024 with approximately 25 m diameter footprints.

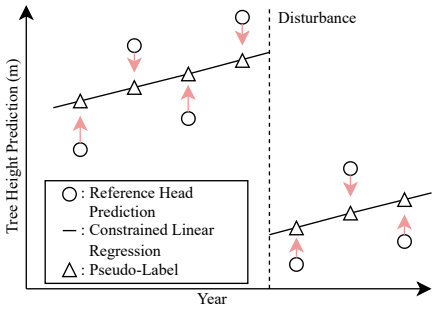


Figure 3: Our growth loss sketch in a disturbance scenario: From the temporally inconsistent predictions of the reference head (circles), we calculate a constrained piecewise linear regression (solid line) to produce pseudo-labels (triangles). The predictions of the prediction head are pushed to better fit the pseudo-labels (red arrows).

Data Processing Pipeline. We create a unified input tensor of shape $18 \times 84 \times 96 \times 96$ (channels $\times$ timesteps $\times$ height $\times$ width) by temporally aligning different input data and spatially resampling all data to 10 m resolution. Quality filtering of the GEDI ground truth ensures reliable measurements.

3.2 Model Architecture

Our model is based on the Swin Transformer architecture (Liu et al., 2021) and combines two key extensions: the Video Swin Transformer (Liu et al., 2022), designed for video input processing, and the Swin-Unet (Cao et al., 2022), tailored for semantic segmentation tasks.

Temporal-Swin-Unet. We combine the architectures from Cao et al. (2022) and Liu et al. (2022) and incorporate some small, yet crucial, modifications to perform pixel-wise regression on a time series of images of shape $C \times T \times H \times W$. We call the resulting architecture *Temporal-Swin-Unet*. It is worth pointing out that, unlike most contemporary approaches, we resort to a patch size of 1×1 pixels, following Nguyen et al. (2025). An illustration is given in Figure 2: The Patch Embed layer linearly projects each voxel[2] into the embedding dimension E. Operating at the original resolution is crucial for our application, where every pixel corresponds to 10×10 meters. The model consists of four Encoder layers and Decoder layers, which are connected via skip connections. Each Encoder and Decoder layer consists of multiple Video Swin Transformer Blocks (Liu et al., 2022). Except at the U-Net's lowest level, all Encoder layers end with a *Temporal Downsample* (TD) layer and all Decoder layers begin with a *Temporal Skip Connection* (TSC).

TD and TSC Layers. The TD layer reduces the temporal and spatial dimensions by applying a year-wise linear projection, concatenating the embeddings of four adjacent pixels, and performing another linear projection to double the embedding size. The TSC layer enriches the Decoder features token-wise per year with the corresponding Encoder features of the same year via a Transformer layer. At the end of the Decoder layer, we upsample spatially by a factor of two.

Conv Heads. On top of the final Decoder layer, we use two heads: the reference head, which is used for pretraining and projects the embeddings voxel-wise to scalar values; and the prediction head, which is added later for fine-tuning and consists of three Conv3D layers with normalization and activation layers in between. Our model thus outputs a tensor of shape $2 \times Y \times H \times W$, representing two canopy height predictions per year and pixel.

3.3 Growth Loss

We propose a self-supervised approach to achieve consistent growth curves, which are monotonically increasing, but allow for sharp cut-offs in disturbance situations, see again Figure 3. Our approach achieves this without affecting single-year accuracy (see our ablations in Appendix A.2.2).

Notation. Let $\boldsymbol{\mathcal{Z}}^{\text{ref}} \in \mathbb{R}^{Y \times H \times W}$ and $\boldsymbol{\mathcal{Z}}^{\text{pred}} \in \mathbb{R}^{Y \times H \times W}$ be the outputs of the reference head and the prediction head. Furthermore, let $\boldsymbol{z}^{\text{ref}} := \boldsymbol{\mathcal{Z}}^{\text{ref}}_{:,h,w} \in \mathbb{R}^{Y}$ and $\boldsymbol{z}^{\text{pred}} := \boldsymbol{\mathcal{Z}}^{\text{ref}}_{:,h,w} \in \mathbb{R}^{Y}$ be the predicted time series at the pixel $(h, w) \in \{1, \dots, H\} \times \{1, \dots, W\}$. In short, the loss fits a regression model on $\boldsymbol{z}^{\text{ref}}$ and uses the fitted values as pseudo-labels for $\boldsymbol{z}^{\text{pred}}$. The regression function is either linear or a combination of two linear functions (pre- and post-disturbance) in case a disturbance is detected, where all slopes are forced to lie in a reasonable interval for tree growth, e.g. in $[s_{\min} = 0\,\text{m/year}, s_{\max} = 3\,\text{m/year}]$.

Disturbance Indicator. We begin by defining what constitutes a disturbance in our context: A disturbance is considered to occur in $\boldsymbol{z}^{\text{ref}} \in \mathbb{R}^{Y}$ when (a) tree height decreased by more than 50% and more than $4\,\text{m}$ and (b) tree height decreased to less than $10\,\text{m}$ within two years.[3] Thus, we define the set of years directly preceeding a disturbance as $\mathbb{Y}_{\text{dstb}}(\boldsymbol{z}^{\text{ref}})$ to be

$$\begin{aligned} \mathbb{Y}_{\text{dstb}}(\boldsymbol{z}^{\text{ref}}) = \{ & y \in \{1, \dots, Y-1\} \mid \\ & \boldsymbol{z}^{\text{ref}}_{y+1} \leq \min(0.5 \cdot \boldsymbol{z}^{\text{ref}}_{y}, \boldsymbol{z}^{\text{ref}}_{y} - 4), \\ & \min(\boldsymbol{z}^{\text{ref}}_{y+1}, \boldsymbol{z}^{\text{ref}}_{y+2}) \leq 10\}. \end{aligned} \tag{1}$$

The local disturbance indicator, defined as the final year preceding a disturbance, is defined as

$$\mathbb{I}_{\text{dstb,loc}}(\boldsymbol{z}^{\text{ref}}) = \min(\mathbb{Y}_{\text{dstb}}(\boldsymbol{z}^{\text{ref}}) \cup \{Y\}) \in \{1, \dots, Y\}^{4}.$$

Combining the pixel-wise local disturbance indicator, we can build an image-wise local disturbance indicator $\mathbb{I}_{\text{dstb,loc}}(\boldsymbol{\mathcal{Z}}^{\text{ref}}) \in \{1, \dots, Y\}^{H \times W}$. Since disturbances generally also influence their surrounding, the final disturbance indicator $\mathbb{I}_{\text{dstb}}(\boldsymbol{\mathcal{Z}}^{\text{ref}}) \in \{1, \dots, Y\}^{H \times W}$ is defined as

$$\mathbb{I}_{\text{dstb}}(\boldsymbol{\mathcal{Z}}^{\text{ref}}) = \text{MinPool}_{3\times3}(\mathbb{I}_{\text{dstb,loc}}(\boldsymbol{\mathcal{Z}}^{\text{ref}})). \tag{2}$$

[2] We define a voxel as a value in the 3D grid $T \times H \times W$, i.e. a pixel at a given timestep.

[3] As no global temporal tree height dataset exists, we define these thresholds solely using domain expertise.

[4] It equals Y if no disturbance is detected; otherwise, it equals the earliest disturbance year. Note that for the majority of tree height prediction time series, there is at most one disturbance year.

Constrained Linear Regression. For some $N \in \mathbb{N}$ and vector $\boldsymbol{z} \in \mathbb{R}^N$, we define the constrained linear regression vector $\hat{\boldsymbol{z}} \in \mathbb{R}^N$ with respect to a minimal and maximal slope $s_{\min} < s_{\max}$ as follows. Let $\tilde{s} \in \mathbb{R}$ be the slope of the simple linear regression model for the dataset $\{(1, \boldsymbol{z}_1), (2, \boldsymbol{z}_2), \ldots, (N, \boldsymbol{z}_N)\}$. Then the slope s, the intercept b and $\hat{\boldsymbol{z}}$ are defined by

$$s \coloneqq \min(\max(\tilde{s}, s_{\min}), s_{\max}) \in [s_{\min}, s_{\max}] \tag{3}$$

$$b \coloneqq \bar{\boldsymbol{z}} - s \cdot \frac{N+1}{2} \in \mathbb{R} \text{ with } \bar{\boldsymbol{z}} \coloneqq \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{z}_n \tag{4}$$

$$\hat{\boldsymbol{z}} \coloneqq s \cdot [1,\ 2\ , \ldots,\ N]^{\mathrm{T}} + b \in R^N. \tag{5}$$

Growth Loss. Pseudo-labels are created by performing piecewise constrained linear regression on the reference time series. Let $y \coloneqq \mathbb{I}_{\mathrm{dstb}}(\boldsymbol{\mathcal{Z}}^{\mathrm{ref}})_{h,w} \in \{1, 2, \ldots, Y\}$ be the detected pre-disturbance year of the reference output and split $\boldsymbol{z}^{\mathrm{ref}}$ into pre- and post-disturbance vectors, that is

$$\boldsymbol{z}^{\mathrm{ref}}_{\mathrm{pre}} \coloneqq [\boldsymbol{z}^{\mathrm{ref}}_1, \ldots, \boldsymbol{z}^{\mathrm{ref}}_y]^{\mathrm{T}} \in \mathbb{R}^y \text{ and } \boldsymbol{z}^{\mathrm{ref}}_{\mathrm{post}} \coloneqq [\boldsymbol{z}^{\mathrm{ref}}_{y+1}, \ldots, \boldsymbol{z}^{\mathrm{ref}}_Y]^{\mathrm{T}} \in \mathbb{R}^{Y-y}.$$

Then the pseudo-labels are defined by concatenating the constrained linear regression vectors $\hat{\boldsymbol{z}^{\mathrm{ref}}_{\mathrm{pre}}}, \hat{\boldsymbol{z}^{\mathrm{ref}}_{\mathrm{post}}}$ for pre- and post-disturbance vectors, thus

$$\hat{\boldsymbol{z}^{\mathrm{ref}}} \coloneqq [\hat{\boldsymbol{z}^{\mathrm{ref}}_{\mathrm{pre}}},\ \hat{\boldsymbol{z}^{\mathrm{ref}}_{\mathrm{post}}}]^{\mathrm{T}} \in \mathbb{R}^Y. \tag{6}$$

Finally, the growth loss measures the distance between pseudo-labels and predictions, i.e.

$$\mathcal{L}_{\mathrm{growth}}(\boldsymbol{z}^{\mathrm{ref}}, \boldsymbol{z}^{\mathrm{pred}}) \coloneqq \frac{1}{Y} ||\hat{\boldsymbol{z}^{\mathrm{ref}}} - \boldsymbol{z}^{\mathrm{pred}}||. \tag{7}$$

3.4 Model Training

Training a spatio-temporal model at a global scale requires careful design of the dataset construction and optimization strategy. The large size and geographic diversity of the input data demand a sampling strategy that balances coverage of relevant forested areas with computational feasibility, both for training and global-scale inference. Further, the sparse and (temporally and spatially) noisy nature of GEDI supervision necessitates specialized training objectives and stable optimization.

Dataset. Building on the multi-sensor, multi-temporal inputs described in Section 3.1, we assembled a large-scale training dataset by sampling spatio-temporal patches centered on GEDI footprints. From each of the 13 000 Sentinel-2 tiles over land with GEDI coverage, we generated up to 230 patches depending on the availability of valid GEDI labels. In non-forested regions with limited relevance (e.g., the Sahara) we restricted the number of patches to a maximum of three to reduce computational overhead and storage requirements. The resulting dataset contains approximately 3 million multi-sensor samples, totaling approx. $50\,\mathrm{TB}$ of input data. For model testing, we selected one hold-out sample per Sentinel-2 tile, ensuring broad spatio-temporal coverage across continents and biomes.

Training Procedure. We trained our model on 8 NVIDIA H200 GPUs for 7 days with the hyperparameters detailed in Table 9 in Appendix A.3. We first pretrained the model using the Huber loss on the reference head for $539\mathrm{k}$ iterations (3 epochs) with a batch size of 16. Subsequent fine-tuning for $179\mathrm{k}$ iterations (0.5 epochs) with a batch size of 8 was performed by training the prediction head with the growth loss detailed in Section 3.3, while freezing the rest of the model parameters. Inference was done on different GPUs (e.g., RTX 4090, A100, H200) taking $80\,000$ GPU hours.

4 Results

We evaluate the *ECHOSAT* map through a comprehensive three-part assessment. Our evaluation uses GEDI labels filtered according to the quality criteria described in Appendix A.1.1, ensuring reliable ground truth. We first assess the prediction accuracy of the *ECHOSAT* map against GEDI labels, then analyze the temporal dynamics and growth patterns captured by our model, and finally compare our predictions against existing single-year baselines (for 2020, see Section 2). When not specified otherwise, reported metrics exclude labels below $5\,\mathrm{m}$ following Hansen et al. (2013a), which defines trees as vegetation exceeding $5\,\mathrm{m}$ in height.

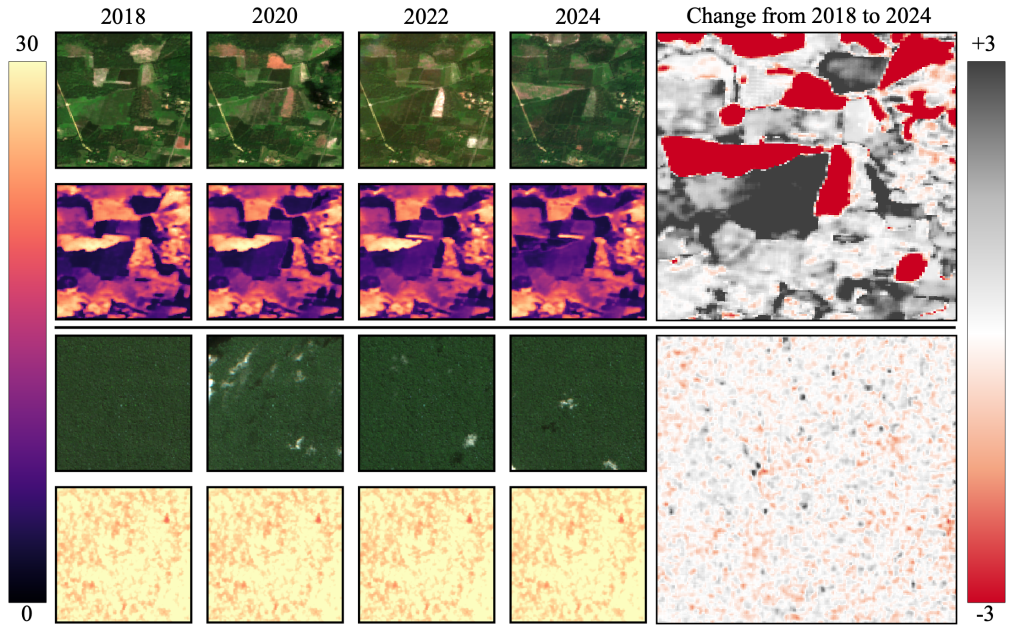


Figure 4: Examples of predicted tree height dynamics for two contrasting regions. Top: Les Landes (France) showing disturbance and regrowth patterns. Bottom: Amazonas (Brazil) with largely stable forest structure. Each block shows optical imagery (top row), predicted tree height (second row), and corresponding change maps from 2018 to 2024 (right column).

4.1 Canopy Height Accuracy

For 2019-2022, yearly MAE values range from $5.36\,\mathrm{m}$ to $6.27\,\mathrm{m}$, indicating similar prediction accuracy across years. However, 2023-2024 show notable variations: MAE increases to $5.79\,\mathrm{m}$ in 2023, then decreases significantly to $4.89\,\mathrm{m}$ in 2024. This pattern correlates with GEDI's operational status, as the instrument was not operational from March 17, 2023, through April 22, 2024, resulting in different label distributions and availability patterns. For details, we refer to Table 6 in Appendix A.2. The substantial gap between MAE ($4.89\,\mathrm{m}-6.27\,\mathrm{m}$) and RMSE ($8.59\,\mathrm{m}-11.21\,\mathrm{m}$) indicates the presence of large prediction errors, suggesting that while most predictions are reasonably accurate, occasional severe errors occur. These errors could stem from remaining noise in the filtered GEDI labels, particularly cases where LiDAR waveforms fail to reach the ground or even to penetrate dense canopies, resulting in ground-level measurements ($0\,\mathrm{m}$) for trees that may be taller than $30\,\mathrm{m}$.

4.2 Canopy Height Growth/Decline

The sparse temporal and spatial distribution of GEDI labels prevents temporal validation, so we assess whether our model captures realistic forest growth patterns. Figure 5 shows median height differences and the lower and upper quartiles for each $1\,\mathrm{m}$ height class from 2018 to each subsequent year for 4 different regions, demonstrating year-to-year growth variations. Pixels are marked as disturbed when the height decreases by more than $5\,\mathrm{m}$ over the time span and are excluded from the median aggregation. The analysis reveals consistent growth across all height classes, with taller trees exhibiting slower growth rates, consistent with established forest growth patterns, and different ecosystems having different growth patterns. Figure 4 shows the predictions and changes for two areas: highly active forests in Les Landes (France) and the Amazonas rainforest (Brazil). In the Les Landes forest in France, which is well known for its intensive wood production and therefore fast-growing tree species, the predictions reveal many disturbances (most likely caused by logging activities) and phases of regrowth. In contrast, the predictions for the Amazonas region remain largely stable, showing very little variation across the different years. Satellite images from 2018 to 2024 together with time series of pixel-wise predictions for five selected pixels around a disturbance in Les Landes are depicted in Figure 9 in Appendix A.2. Please note that our model is able to predict canopy height even for the first year, 2018, when GEDI labels are not available.

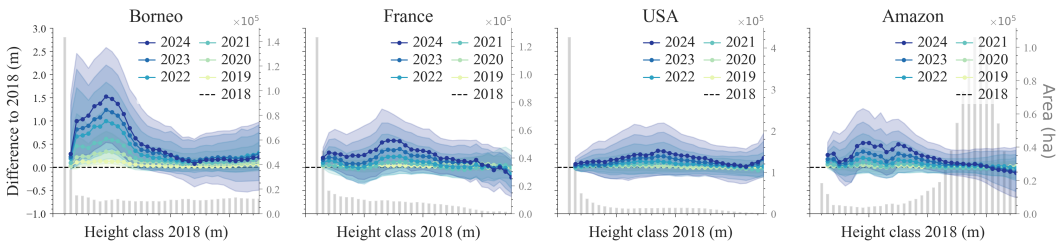


Figure 5: Median height difference (excluding disturbances with more than 5m height decrease) from 2018 to each year, binned in $1\,\mathrm{m}$ height classes for Borneo, USA, France and Amazon. The right y-axis shows the height class distribution and area for these classes.

Table 1: Comparison of global-scale methods for 2020 regarding MAE (m), MSE (m^2), RMSE (m), MAPE (%), coefficient of determination R^2 and Pearson correlation coefficient r, with R^2_{all} and r_{all} computed on all labels, including labels below $5\,\mathrm{m}$.

Method	MAE ↓	MSE ↓	RMSE ↓	MAPE ↓	R^2 ↑	R^2_{all} ↑	r ↑	r_{all} ↑
Potapov et al. (2021)	9.11	185.52	13.62	54.90	−0.20	0.31	0.50	0.70
Lang et al. (2023)	7.97	143.78	11.99	53.33	0.07	0.43	0.52	0.71
Pauls et al. (2024)	6.85	138.19	11.76	34.20	0.10	0.50	0.51	0.73
Tolan et al. (2024)	11.89	260.28	16.13	68.78	−0.69	0.04	0.45	0.64
Ours	**5.85**	**118.07**	**10.87**	**30.20**	**0.23**	**0.58**	**0.59**	**0.77**

4.3 Comparison Against Existing Maps

While no global-scale temporal tree height maps exist, four single-year approaches provide suitable baselines for a single-year comparison: Tolan et al. (2024) (DINOv2-based, $1\,\mathrm{m}$ resolution), Potapov et al. (2021) (Random Forest, $30\,\mathrm{m}$ Landsat), Lang et al. (2023) (CNN, $10\,\mathrm{m}$ Sentinel-2), and Pauls et al. (2024) (U-Net, $10\,\mathrm{m}$ Sentinel-2). We compare our 2020 predictions against these baselines using the same MAE, MSE, RMSE, MAPE, R^2 and Pearson correlation r on our test samples. All baseline maps were downloaded from Google Earth Engine, rescaled to $10\,\mathrm{m}$ using bilinear interpolation, and warped to the corresponding Sentinel-2 tile CRS. Table 1 reports the metrics, showing that *ECHOSAT* consistently outperforms the other maps, while also reducing the variance in its predictions. The same ranking holds on independent airborne laser scanning data from NEON sites in the United States (Appendix A.2.6). Scatterplots against these labels resemble the GEDI scatterplots, with substantially less noise about the diagonal, which indicates that the gain is not only a closer fit to GEDI noise.

Figure 1 shows a visual comparison of all maps in 3 distinct regions. Although the map by Tolan et al. (2024) is resampled to $10\,\mathrm{m}$, it still shows fine details is well suited for the detection of smaller tree patches. The map by Potapov et al. (2021) uses $30\,\mathrm{m}$ Landsat data as input and therefore fails to identify some trees; however, its accuracy is higher than that of the map of Tolan et al. (2024). The models by Lang et al. (2023) and Pauls et al. (2024) and our model use Sentinel-2 as input and can detect most smaller forest patches, while also achieving higher accuracy against the reference labels. Further, the model by Pauls et al. (2024) and our model show finer structure in their predictions than the model by Lang et al. (2023). The scatterplots in Figure 6 reveal that the predictions by Tolan et al. (2024), Pauls et al. (2024) and Potapov et al. (2021) saturate between $30\,\mathrm{m}$ and $35\,\mathrm{m}$, while those by Lang et al. (2023) and our map extend beyond this range. As already indicated by the correlation coefficient, the body of our scatterplot is also narrower than that ofLang et al. (2023). Further figures are provided in Appendix A.2.

5 Conclusion

Our approach addresses the fundamental limitation of existing static forest height products through an adapted growth loss framework that inherently enforces physically plausible forest dynamics without requiring post-processing and without decreasing single-year accuracy. By leveraging multi-sensor satellite data, we demonstrate how temporal forest monitoring can be achieved at an unprecedented

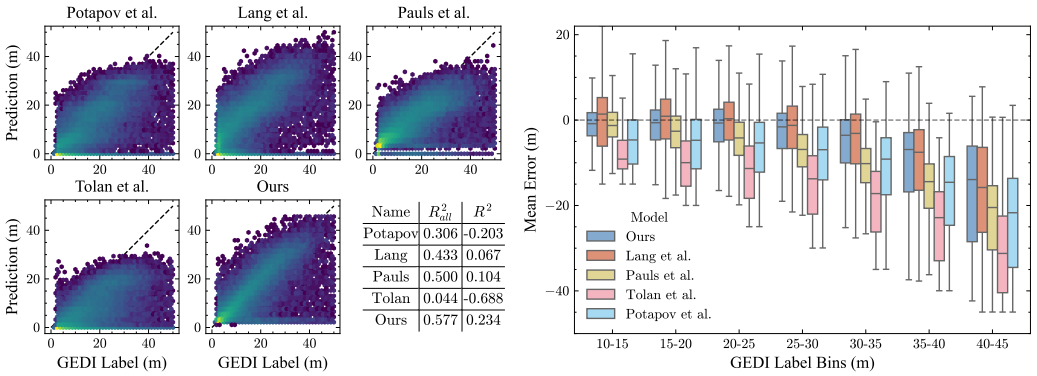


Figure 6: Left: Scatterplots showing the predicted height for 2020 vs. GEDI labels and a table comparing the Pearson correlation coefficients for all labels (r_{all}) and for labels exceeding $5\,\text{m}$ (r). Right: Error distribution analysis across height classes ($5\,\text{m}$ bins) for all baseline methods. Boxplots show mean absolute error for each height class, revealing that tall tree prediction remains challenging across all approaches, with errors increasing substantially for heights above $25\,\text{m}$.

scale and resolution. Our evaluation indicates that different height classes of trees have varying growth rates, consistent with existing literature, with trees between $10\,\text{m}$ and $20\,\text{m}$ having the highest average growth rate. In single-year comparisons with existing maps, *ECHOSAT* shows improved accuracy across all evaluated metrics, both against held-out GEDI labels and against independent NEON airborne laser scanning data. Overall, we believe our work advances global forest monitoring for climate change mitigation, carbon accounting, and disturbance assessment.[5]

5.1 Limitations

Smaller annual growth increments often fall within model and data uncertainty, making small height changes difficult to detect reliably. Our disturbance criteria effectively capture clear-cuts and major fires but may miss gradual degradation, selective logging, or scattered tree mortality. Furthermore, the proposed growth loss is based on a few hyperparameters that need to be manually set, such as the maximum allowed growth per year or the decrease in tree height to classify a change as a disturbance. These parameters were defined in close collaboration with forest experts, but may evolve with increasing knowledge and can vary across experts. Finally, GEDI's sparse temporal coverage does not allow direct validation of year-to-year height changes. While Section 4.2 shows that our predictions follow realistic ecological growth patterns, validation against independent repeated observations (airborne LiDAR campaigns, permanent forest plots) is needed. The seven-year timespan (2018–2024) also limits the assessment of long-term dynamics and multi-decadal carbon accumulation patterns.

5.2 Broader Impact

Accurate mapping of tree height is a prerequisite for assessing biomass carbon storage and wood resources accross the world's forests. For monitoring forest changes under human and climate pressure, we need dynamic maps instead of static products. Furthermore, as most height loss instances come from small scale events such as mortality occurring in clusters of trees, natural forest disturbances, and human activities including timber harvest, degradation and deforestation, a high spatial resolution is needed to capture the fine-scale patterns of height decreases. The new global annual forest height maps at $10\,\text{m}$ resolution developed in this study with a model that can learn to reconstruct height changes not only from spatial gradients but also from temporal images represent a significant step forward that goes beyond previous global static maps and extends temporal change maps that were limited to a few regions and used coarser-resolution models. Our map was evaluated against held-out GEDI labels and, independently over a limited area, against ALS data in the USA (Appendix A.2.6). For pixels not affected by losses where forests are growing or regrowing after previous loss events, we also showed regular year-on-year height increments that are consistent with ecological knowledge indicating that younger and shorter trees grow faster than taller ones. Hence, our maps can be used for a variety of forest management tasks.

[5]The produced maps will be made available upon acceptance.

AI USE STATEMENT

In this work, we used generative AI tools to check the grammar and spelling of our text. We further used AI for the coding tasks associated with model development, inference and ablation studies. We did not use any large language models (LLMs) to brainstorm research ideas or write entire paragraphs. All LLM use was non-agentic (i.e., without loops or goal-directed execution) We take responsibility for the final content of this work, including all text, claims and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We document the full experimental pipeline in the main paper and appendix. Data sources, preprocessing, resampling, normalization, and GEDI quality filters are described in Section 3.1 and Appendix A.1.1 (Table 2). Dataset construction, the per-tile hold-out split, and the spatial-autocorrelation constraint are specified in Section 3.4 and Appendix A.1.2. The Temporal-Swin-Unet architecture, including the Temporal Downsample and Temporal Skip Connection layers and the two convolutional heads, is defined in Section 3.2 and Appendix A.1.3. The growth loss, disturbance indicator, and constrained piecewise linear regression are fully specified in Section 3.3. Training stages, hardware, and hyperparameters are reported in Section 3.4 and Table 9. Evaluation protocols, baseline comparisons, and growth-loss ablations appear in Section 4 and Appendix A.2. Predicted maps can be inspected at `https://echosatanonymous.projects.earthengine.app/view/echosat-temporal` and will be released upon acceptance. The code for training is supplied in the supplementary material.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: One earth observation model for many resolutions, scales, and modalities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19530–19540, 2025.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Ralph Dubayah, John Armston, Sean P. Healey, Jamis M. Bruening, Paul L. Patterson, James R. Kellner, Laura Duncanson, Svetlana Saarela, Göran Ståhl, Zhiqiang Yang, Hao Tang, J. Bryan Blair, Lola Fatoyinbo, Scott Goetz, Steven Hancock, Matthew Hansen, Michelle Hofton, George Hurtt, and Scott Luthcke. GEDI launches a new era of biomass inference from space. *Environmental Research Letters*, 17(9):095001, August 2022. ISSN 1748-9326. doi: 10.1088/1748-9326/ac8694.

Samuel Favrichon, Jake Lee, Yan Yang, Ricardo Dalagnol, Fabien Wagner, Le Bienfaiteur Sagang, and Sassan Saatchi. Monitoring changes of forest height in California. *Frontiers in Remote Sensing*, 5, January 2025. ISSN 2673-6187. doi: 10.3389/frsen.2024.1459524.

Ibrahim Fayad, Max Zimmer, Martin Schwartz, Fabian Gieseke, Philippe Ciais, Gabriel Belouze, Sarah Brood, Aurelien De Truchis, and Alexandre d'Aspremont. Dunia: Pixel-sized embeddings via cross-modal alignment for earth observation applications. *arXiv preprint arXiv:2502.17066*, 2025.

Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023.

M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853, 2013a. doi: 10.1126/science.1244693. URL https://www.science.org/doi/abs/10.1126/science.1244693.

M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160):850–853, November 2013b. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1244693.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

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.

Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the Earth. *Nature Ecology & Evolution*, pp. 1–12, September 2023. ISSN 2397-334X. doi: 10.1038/s41559-023-02206-6.

Siyu Liu, Martin Brandt, Thomas Nord-Larsen, Jerome Chave, Florian Reiner, Nico Lang, Xiaoye Tong, Philippe Ciais, Christian Igel, Adrian Pascual, Juan Guerra-Hernandez, Sizhuo Li, Maurice Mugabowindekwe, Sassan Saatchi, Yuemin Yue, Zhengchao Chen, and Rasmus Fensholt. The overlooked contribution of trees outside forests to tree cover and woody biomass across Europe. *Science Advances*, 9(37):eadh4097, September 2023. doi: 10.1126/sciadv.adh4097.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

National Ecological Observatory Network (NEON). Discrete return lidar point cloud (dp1.30003.001), 2026. URL https://data.neonscience.org/data-products/DP1.30003.001/RELEASE-2026.

Duy Kien Nguyen, Mido Assran, Unnat Jain, Martin R. Oswald, Cees G. M. Snoek, and Xinlei Chen. An image is worth more than 16x16 patches: Exploring transformers on individual pixels. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tjNf0L8QjR.

Yude Pan, Richard A. Birdsey, Oliver L. Phillips, Richard A. Houghton, Jingyun Fang, Pekka E. Kauppi, Heather Keith, Werner A. Kurz, Akihiko Ito, Simon L. Lewis, Gert-Jan Nabuurs, Anatoly Shvidenko, Shoji Hashimoto, Bas Lerink, Dmitry Schepaschenko, Andrea Castanho, and Daniel Murdiyarso. The enduring world forest carbon sink. *Nature*, 631(8021):563–569, July 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07602-x.

Jan Pauls, Max Zimmer, Una M. Kelly, Martin Schwartz, Sassan Saatchi, Philippe Ciais, Sebastian Pokutta, Martin Brandt, and Fabian Gieseke. Estimating Canopy Height at Scale, June 2024.

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 00344257. doi: 10.1016/j.rse.2020.112165.

Johannes Reiche, Adugna Mullissa, Bart Slagter, Yaqing Gou, Nandin-Erdene Tsendbazar, Christelle Odongo-Braun, Andreas Vollrath, Mikaela J. Weisse, Fred Stolle, Amy Pickens, Gennadii Donchyts, Nicholas Clinton, Noel Gorelick, and Martin Herold. Forest disturbance alerts for the Congo Basin using Sentinel-1. *Environmental Research Letters*, 16(2):024005, January 2021. ISSN 1748-9326. doi: 10.1088/1748-9326/abd0a8.

Martin Schwartz, Philippe Ciais, Aurélien De Truchis, Jérôme Chave, Catherine Ottlé, Cedric Vega, Jean-Pierre Wigneron, Manuel Nicolas, Sami Jouaber, Siyu Liu, Martin Brandt, and Ibrahim Fayad. FORMS: Forest Multiple Source height, wood volume, and biomass maps in France at 10 to 30 m resolution based on Sentinel-1, Sentinel-2, and Global Ecosystem Dynamics Investigation (GEDI) data with a deep learning approach. *Earth System Science Data*, 15(11): 4927–4945, November 2023. ISSN 1866-3508. doi: 10.5194/essd-15-4927-2023.

Martin Schwartz, Philippe Ciais, Ewan Sean, Aurélien De Truchis, Cédric Vega, Nikola Besic, Ibrahim Fayad, Jean-Pierre Wigneron, Sarah Brood, Agnès Pelissier-Tanon, Jan Pauls, Gabriel Belouze, and Yidi Xu. Retrieving yearly forest growth from satellite data: A deep learning based approach. *Remote Sensing of Environment*, 330:114959, December 2025. ISSN 00344257. doi: 10.1016/j.rse.2025.114959.

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.

Jamie Tolan, Hung-I Yang, Benjamin Nosarzewski, Guillaume Couairon, Huy V. Vo, John Brandt, Justine Spore, Sayantan Majumdar, Daniel Haziza, Janaki Vamaraju, Theo Moutakanni, Piotr Bojanowski, Tracy Johns, Brian White, Tobias Tiecke, and Camille Couprie. Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment*, 300:113888, January 2024. ISSN 0034-4257. doi: 10.1016/j.rse.2023.113888.

Erkki Tomppo, Håkan Olsson, Håkan Olsson, Göran Ståhl, Mats Nilsson, Olle Hagner, and Matti Katila. Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sensing of Environment*, 112(5):1982–1999, May 2008. doi: 10.1016/j.rse.2007.03.032.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. *arXiv preprint arXiv:2502.09356*, 2025.

Svetlana Turubanova, Peter Potapov, Matthew C. Hansen, Xinyuan Li, Alexandra Tyukavina, Amy H. Pickens, Andres Hernandez-Serna, Adrian Pascual Arranz, Juan Guerra-Hernandez, Cornelius Senf, Tuomas Häme, Ruben Valbuena, Lars Eklundh, Olga Brovkina, Barbora Navrátilová, Jan Novotný, Nancy Harris, and Fred Stolle. Tree canopy extent and height change in Europe, 2001–2021, quantified using Landsat data archive. *Remote Sensing of Environment*, 298:113797, December 2023. ISSN 0034-4257. doi: 10.1016/j.rse.2023.113797.

Yuxin Wu and Kaiming He. Group normalization. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, pp. 3–19, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01260-1. doi: 10.1007/978-3-030-01261-8_1. URL https://doi.org/10.1007/978-3-030-01261-8_1.

A Appendix

A.1 Methodology

A.1.1 Data

Here we describe data sources used and their processing in more detail. Table 2 provides an overview.

Sentinel-2 Optical Data. We use all 12 spectral bands from Sentinel-2 L2A products, which provide atmospheric correction and cloud probability estimates. For each year, we select one image per calendar month based on the highest percentage of valid pixels (excluding cloudy and black pixels as identified by the Sen2Cor algorithm). The Sentinel-2 $10\,\mathrm{m}$ bands serve as the foundation for our dataset, hence bands with $20\,\mathrm{m}$ and $60\,\mathrm{m}$ native resolution are upsampled to $10\,\mathrm{m}$ using nearest neighbor interpolation. Values are normalized to the range [-1, +1] using band-specific scaling factors: bands 1-4 are scaled from [0, 2000], bands 6-9 from [0, 6000], band 0 from [0, 1000], and bands 5, 10-11 from [0, 4000].

Sentinel-1 Radar Data. We utilize C-band synthetic aperture radar data with $10\,\mathrm{m}$ spatial resolution, processing quarterly median composites of VH polarization for both ascending and descending orbits. Digital numbers are converted to backscatter coefficients (dB) and scaled from [-50, +1] to [-1, +1]. The downloaded images from Google Earth Engine are already geospatially aligned to the $10\,\mathrm{m}$ Sentinel-2 bands and can therefore be stacked directly. To align with monthly Sentinel-2 data, each quarterly composite is duplicated across the corresponding three months.

ALOS PALSAR-2 Radar Data. We incorporate L-band synthetic aperture radar data with $25\,\mathrm{m}$ spatial resolution, using yearly median composites of HH and HV polarizations. Digital numbers are converted to backscatter coefficients (dB) and scaled from [-50, +1] to [-1, +1]. After reprojection and upsampling using bilinear interpolation with the pixels being aligned to the Sentinel-2 $10\,\mathrm{m}$ bands (using `gdalwarp`'s `-tap` option), the yearly composite is duplicated across all 12 months to maintain temporal consistency.

TanDem-X Data. We utilize two products from the TanDEM-X mission: (1) a $30\,\mathrm{m}$ resolution digital elevation model scaled from [0, 7000 m] to [-1, +1], and (2) a forest/non-forest classification map with 3 classes normalized to [-1, +1]. Both products are duplicated across all 84 time steps (7 years × 12 months) as they represent static features.

GEDI LiDAR Ground Truth. We use GEDI L2A V2 products as ground-truth labels, applying quality filters to ensure data reliability: relative height at the 98th percentile (rh98) between $0\,\mathrm{m}$ and $150\,\mathrm{m}$, only high-power beams, number of detected modes $>= 1$, quality flag = 1, degrade flag = 0, and sensitivity $>= 0.95$. These measurements provide sparse tree height estimates with approximately $25\,\mathrm{m}$ diameter footprints. Labels are rasterized to the Sentinel-2 $10\,\mathrm{m}$ bands using the provided geolocation of the highest return (given by `lon_highestreturn` and `lat_highestreturn`). In the very rare cases where multiple GEDI labels correspond to the same pixel and year associated, the maximum is taken. Although the GEDI diameter is often referred to as $25\,\mathrm{m}$, we use only the center pixel, as the returned energy decreases rapidly with distance from the center (see Figure 8).

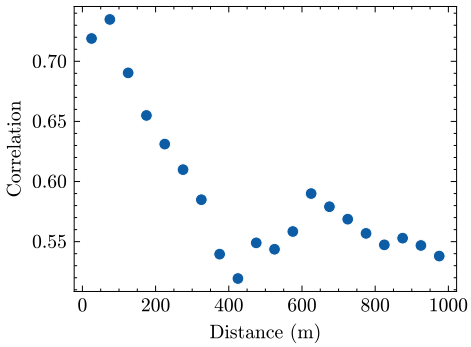
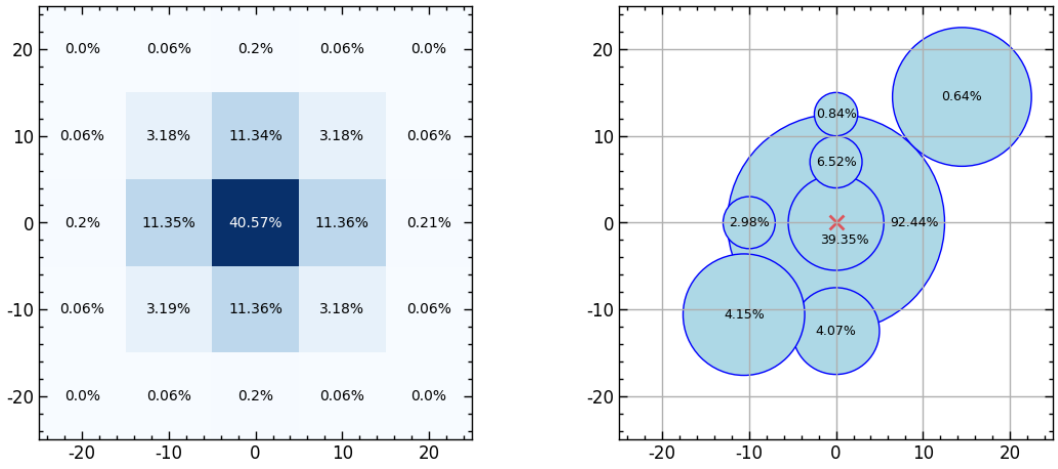


Figure 7: Analysis of spatial autocorrelation between GEDI labels in our dataset. Correlation is given by the Pearson product-moment correlation coefficient in bins of $50\,\mathrm{m}$. The correlation starts to level off at around $300\,\mathrm{m}$ to $400\,\mathrm{m}$.

To ensure training data quality and focus on forested areas, we apply additional spatial filtering using TanDem-X data. We calculate terrain slope within a $70\,\mathrm{m}$ radius around each GEDI measurement and exclude locations with slopes exceeding $20°$ to avoid bare mountain areas. Additionally, we use the TanDem-X Global Urban Footprint to remove measurements where the human footprint exceeds 10%, ensuring our model trains on natural forest environments rather than urban areas.



(a) (b)

Figure 8: GEDI shot photon distribution. (a) The photon density of a single GEDI shot follows a normal distribution across both spatial dimensions. A Sentinel-2 pixel located directly at the center of the footprint receives 40.57% of the emitted laser energy. (b) Percentage of emitted photons received by sample trees at different positions. A tree with a $5\,\mathrm{m}$ radius centered directly under the GEDI footprint receives approximately 40% of the photons, while a tree with a $7\,\mathrm{m}$ radius offset by $15\,\mathrm{m}$ from the center receives only 4.15%. Note that these values represent the emitted photon distribution, not the reflected energy measured by the sensor. Actual measurements depend on material reflectance properties, ground return proportion, and the `rh98` metric extraction, requiring substantially more than 2% photon coverage for a tree to be counted as the tallest tree in a GEDI measurement.

Table 2: Data sources used in *ECHOSAT* with their specifications and preprocessing parameters.

Source	Origin	Res.	CRS	Resampling	Bands/Channels	Normalization
Sentinel-1	Google Earth Engine	10m	UTM	NN	VH (Asc/Des)	[-50, 1] in dB
Sentinel-2	AWS	10m	UTM	N/A	All w/o B10	[0, 1000]: 0 [0, 2000]: 1–4 [0, 6000]: 6–9 [0, 4000]: 5/10/11
ALOS Palsar-2	JAXA FTP	25m	EPSG:4326	Bilinear	HH/HV	[-50, 1] in dB
TandemX EDEM	DLR Geoservice	30m	EPSG:4326	Bilinear	—	[0, 7000]
TandemX FNF	DLR Geoservice	50m	EPSG:4326	NN	—	[0–2]
GEDI L2A	NASA Earthdata	∼25m	EPSG:4326	N/A	rh98	—

A.1.2 Train/Test-Split

Our testing dataset is created by randomly picking one area ($960\,\mathrm{m} \times 960\,\mathrm{m}$) inside each Sentinel-2 tile. To limit the effect of spatial autocorrelation, we enforce a minimum distance between training and testing patches of at least $360\,\mathrm{m}$. Figure 7 shows an analysis of spatial autocorrelation for our dataset, where the correlation is given by the Pearson product-moment correlation coefficient. The correlation starts at ≈ 0.74 and levels off at approximately 0.55 at a lag between $300\,\mathrm{m}$ and $400\,\mathrm{m}$.

A.1.3 Model Architecture

Here, we define the model architecture and design decisions in more detail.

Patch Embed. We implement the Patch Embed via a Conv3D layer, with kernel size and stride set to $(1, 1, 1)$, which is equivalent to applying a linear layer channel-wise for every pixel and every timestep. Embedding patches of size 4×4 pixels, as most other contemporary approaches do, would lead to the model producing blurry forest borders and overlooking individual or extraordinarily tall trees.

Table 3: Distribution of year-to-year growth (m) predicted by the reference and prediction heads across growth bins for all pixels and disturbance pixels. The prediction head tends to concentrate mass in the $(0.0, 1.0]$ bin, underrepresenting moderate negative and extreme growth categories.

	All pixels		Disturbance pixels	
Growth Bin	Reference Head	Prediction Head	Reference Head	Prediction Head
$(-\infty,\ -4.0]$	0.25%	0.15%	19.05%	17.17%
$(-4.0,\ -0.2]$	15.89%	4.06%	23.37%	13.09%
$(-0.2,\ 0.0]$	37.12%	31.82%	14.02%	8.34%
$(0.0,\ 1.0]$	44.35%	63.42%	28.44%	42.08%
$(1.0,\ 3.0]$	2.18%	0.53%	9.36%	17.02%
$(3.0,\ \infty)$	0.22%	0.03%	5.75%	2.29%

Temporal Downsample (TD) Layer. The TD layer takes as input a tensor of shape $T_{\text{in}} \times H_{\text{in}} \times W_{\text{in}} \times E_{\text{in}}$ and outputs a tensor of shape `reduce_time[`l_{enc}`]` $\times H_{\text{in}}/2 \times W_{\text{in}}/2 \times 2E_{\text{in}}$. In our model, we set `reduce_time = [28, 14, 7]`, so the time dimension is reduced from 84 to 28 by a factor of three in Encoder layer 1, and then twice by a factor of two. The temporal reduction is implemented via a linear layer (without a bias) applied individually for every year and pixel by concatenating all embeddings of a pixel for a given year, then linearly projecting the vector down to the target temporal resolution. Afterwards, the spatial resolution is halved by concatenating the embeddings of four spatially adjacent pixels, applying Layer Normalization (Ba et al., 2016) and then applying another linear projection. The layer's output is the input for the following Encoder layer, and for the TSC layer of the corresponding Decoder layer via a skip connection.

Temporal Skip Connection (TSC) Layer. The TSC layer is the Decoder counterpart of the TD layer. Note how the time dimension changes throughout the model: it is iteratively reduced from 84 to 7 in the Encoder, but stays 7 throughout the whole Decoder, as we need a single prediction map per year. This prohibits the simple addition of Encoder and Decoder inputs in the TSC layer. After trying out multiple designs, we settled on a Transformer layer, which we will now explain in detail for the skip connection between Encoder layer 1 and Decoder layer 4. The input coming from Encoder layer 1 has shape $84 \times H \times W \times E$ and the output of Decoder layer 4 has shape $7 \times H \times W \times E$. Now, we reshape and concatenate these inputs into a tensor of shape $7HW \times \frac{84+7}{7} \times E$. For every pixel and year (= voxel), we have 13 features, one from the decoder, the rest from the encoder. The decoder feature of a pixel can thus attend to its encoder features of the same year, before being passed on. After the attention we only keep the decoder token and ignore the others.

3D Window Multi-Head Self Attention (3D W-MSA). In the 3D window multi-head self attention, each token can attend to the other tokens within the same window, which spans two tokens along the temporal dimension and six along both spatial dimensions. In typical Video Swin Transformer fashion, every other attention block is shifted in time and space dimension. Thus, every token can attend to the 71 other tokens in the same window, 36 of which are from the previous or following timestep. Using an efficient attention mechanism is necessary due to the otherwise quadratic complexity in the number of pixels HW in an image. Windowed attention is a strong contender, inducing a locality bias while sacrificing the global receptive field in return. In subsequent Encoder layers, the receptive field of the windowed attention is progressively doubled, due to the downsampling operations.

Conv Heads. The reference head consists of a single 3D convolution that projects linearly from the embedding dimension to a scalar per voxel. The prediction head starts with two 3D convolutions with kernel size 3 in the temporal and spatial dimensions, keeping the embedding dimension unchanged, followed by Group Normalization (Wu & He, 2018) and ReLU activation. Therefore, neighboring pixels and consecutive years are able to interact with each other. The final step is a single 3D convolution as in the reference head. The prediction head is designed to allow local spatial and temporal interaction, in order to facilitate the precise detection of forest borders or disturbance locations. Without this interaction, our models tend to have problems with border detection, presumably due to geolocation uncertainty of GEDI measurements.

Table 4: Ablation: Comparison of our approach with fine-tuning (using GrowthLoss or simple linear regression) and without fine-tuning for 2020 labels above $5\,\mathrm{m}$ in terms of MAE (m), MSE (m^2), RMSE (m), MAPE (%), and Pearson correlation coefficients r and r_{all} (the latter computed on all labels, including labels below $5\,\mathrm{m}$). The IQR is given in square brackets. Fine-tuning on pseudo-labels with our proposed GrowthLoss (see Section 3.3) does not substantially affect the accuracy against the reference labels.

Method	MAE ↓	MSE ↓	RMSE ↓	MAPE ↓	r ↑	r_{all} ↑
Ours (w/ GrowthLoss)	5.85 [4.90]	**118.07** [36.16]	**10.87** [4.90]	30.20 [33.32]	**0.59**	**0.77**
Ours (w/ Linear)	5.87 [4.89]	118.51 [36.14]	10.89 [4.89]	30.34 [33.36]	**0.59**	**0.77**
Ours (w/o fine-tuning)	**5.84** [4.89]	119.36 [35.55]	10.93 [4.89]	**29.44** [32.13]	**0.59**	**0.77**

A.2 Results

A.2.1 Ablation: Effect of Growth Loss on Single-Year Accuracy

We present an ablation study of our fine-tuning methodology, comparing the model without fine-tuning, with fine-tuning using our GrowthLoss, and with fine-tuning using simple linear regression. Table 4 shows that all three approaches yield similar results in terms of quantitative metrics, which is expected since fine-tuning only uses the pseudo-labels produced by the pretrained backbone. The impact of fine-tuning becomes evident when examining the temporal evolution of pixel predictions (Figure 9). Without fine-tuning, the model can capture both tree growth and cutting events, but exhibits unwanted height fluctuations. Linear regression fine-tuning smooths tree growth predictions but fails to properly capture disturbances. Fine-tuning with our GrowthLoss, in contrast, reduces uncertainty in growth pixels, detects disturbances more accurately, and minimizes height fluctuations in disturbance border areas.

A.2.2 Ablation: Effect of Growth Loss on Temporal Consistency

Here we do an ablation study showing the effect of the proposed Growth Loss on year-to-year fluctuations compared to a standard Huber Loss. Table 3 reports the relative frequency of year-to-year height differences (i.e., growth = $\hat{h}_{t+1} - \hat{h}_t$) falling into several ecologically meaningful bins. We compare the *reference head*, trained without the growth loss, to the *prediction head*, trained with the proposed growth loss. To obtain these statistics, we sampled 3,000,000 pixels and computed the distribution of growth values both across (1) all pixels and (2) the subset of pixels where a disturbance event has been identified. Several key observations can be made:

Reduction of spurious negative growth. Across all pixels, the prediction head substantially reduces the proportion of predictions falling in the $(-4.0,\, -0.2]\,\mathrm{m}$ bin from 15.89% to 4.06%. This bin largely captures implausible year-to-year height decreases in undisturbed forest—precisely the type of temporal inconsistency the growth loss is designed to suppress. Height decreases of this magnitude are ecologically unrealistic in the absence of disturbance events, and their reduction indicates that the growth loss successfully discourages physically implausible predictions.

Concentration in the ecologically realistic growth range. The prediction head shifts the majority of predictions into the $(0.0,\, 1.0]\,\mathrm{m}$ bin (from 44.35% for the reference head to 63.42% for the prediction head). This range is consistent with expected annual canopy height growth rates in temperate and boreal forests, suggesting that the growth loss encourages the model to produce temporally smoother and more ecologically plausible trajectories.

Plausible disturbance dynamics. When restricting the analysis to pixels with identified disturbances, each pixel is expected to exhibit roughly one disturbance event across six consecutive growth periods (corresponding to seven prediction years), yielding an expected disturbance frequency of approximately $\frac{1}{6} \approx 17\%$. The prediction head brings the proportion of strong negative growth in the $(-\infty,\, -4.0]\,\mathrm{m}$ bin closer to this expected rate (17.17%, compared to 19.05% for the reference head), while simultaneously reducing moderate spurious losses in the $(-4.0,\, -0.2]\,\mathrm{m}$ bin from 23.37%

Table 5: Test-time data-source ablation on a sample of the GEDI test set. Each row keeps the listed sensors and zeros the remaining input channels. S2 is Sentinel-2, S1 is Sentinel-1, PALSAR-2 is ALOS PALSAR-2, and DEM is the TanDEM-X elevation model. "All" adds the TanDEM-X forest/non-forest map. MAE (m), MSE (m^2), RMSE (m), MAPE (%), R^2 and Pearson correlation r, with R^2_{all} and r_{all} computed on all labels, including labels below $5\,\mathrm{m}$.

Configuration	MAE ↓	MSE ↓	RMSE ↓	MAPE ↓	R^2 ↑	R^2_{all} ↑	r ↑	r_{all} ↑
S2	7.045	143.153	11.965	34.526	0.071	0.477	0.522	0.709
S2 + S1	6.653	134.027	11.577	32.418	0.131	0.511	0.542	0.731
S2 + S1 + PALSAR-2	6.613	134.879	11.614	31.964	0.125	0.516	0.549	0.745
S2 + S1 + PALSAR-2 + DEM	5.823	118.903	10.904	**29.151**	0.229	0.573	0.588	**0.771**
All	**5.788**	**117.961**	**10.861**	29.293	**0.235**	**0.577**	**0.591**	**0.771**

to 13.09%. This indicates that the growth loss preserves the model's ability to capture genuine disturbance-driven canopy loss while suppressing artifactual negative growth.

Rather than viewing the growth loss purely through the lens of per-timestep accuracy, we emphasize that it achieves its intended purpose: improving temporal consistency without degrading performance.

A.2.3 Ablation: Data Sources

Retraining the model for each sensor subset is computationally not feasible, so we report a test-time input ablation. Our trained model is left unchanged and we zero out the channels for one sensor at inference. Metrics in Table 5 are computed on a sample of the test set. Because the model was trained with all sensors present, zeroing a group at test time is a harsher intervention than retraining without that group.

A.2.4 Year-Wise Evaluation

Table 6 presents MAE, MSE, RMSE, MAPE, R^2 and Pearson correlation r for our *ECHOSAT* maps for each year individually. Errors generally decrease over time, which may suggest that recent years are easier to predict, but could also reflect improvements in GEDI label processing by NASA. A similar pattern is observed when examining errors across different height bins for each year (Figure 10).

Table 6: Year-wise comparison of MAE (m), MSE (m^2), RMSE (m), MAPE (%), coefficient of determination R^2 and Pearson correlation coefficient r, with R^2_{all} and r_{all} computed on all labels, including labels below $5\,\mathrm{m}$.

Year	MAE ↓	MSE ↓	RMSE ↓	MAPE ↓	R^2 ↑	R^2_{all} ↑	r ↑	r_{all} ↑
2019	6.09	123.44	11.11	30.29	0.23	0.58	0.59	0.77
2020	5.85	118.07	10.87	30.20	0.23	0.58	0.59	0.77
2021	5.41	96.03	9.80	29.87	0.29	0.62	0.63	0.79
2022	5.24	85.37	9.24	29.48	0.35	0.66	0.66	0.82
2023	5.56	102.86	10.14	29.31	0.30	0.62	0.62	0.79
2024	4.77	72.01	8.49	27.54	0.39	0.68	0.68	0.83

A.2.5 Spatial Distribution of Errors

Figure 11 (a) shows a global map where the average error in each region is indicated by color (darker colors correspond to higher errors). Because height and error are closely correlated (as seen in Table 7 for labels $< 5\,\mathrm{m}$), this map primarily reflects the average tree height per region. Figures 11 (b–g) show the same map broken down into $5\,\mathrm{m}$ height bins. While errors generally increase with height, the relatively small errors for large trees in the Amazonas rainforest and Congo Basin may indicate lower structural heterogeneity in these forests, making them easier to predict.

Table 7: Comparison of global-scale methods for 2020 only for labels below $5\,\mathrm{m}$ regarding MAE (m), MSE (m^2), RMSE (m), MAPE (%), Pearson correlation coefficient r and r_{all} (on all labels, including labels above $5\,\mathrm{m}$). Please note that while GEDI measures ground elevation between $2\,\mathrm{m}$ and $3\,\mathrm{m}$, Tolan et al. (2024); Lang et al. (2023); Potapov et al. (2021) either use other labels or manually set their ground predictions to $0\,\mathrm{m}$, thereby making a comparison here ineffective.

Method	MAE ↓	MSE ↓	RMSE ↓	MAPE ↓	r ↑	r_{all} ↑
Potapov et al. (2021)	2.84	9.08	3.01	98.20	0.26	0.70
Lang et al. (2023)	3.04	13.30	3.65	104.08	0.28	0.71
Pauls et al. (2024)	1.08	2.99	1.73	37.04	**0.49**	0.73
Tolan et al. (2024)	2.87	8.93	2.99	99.04	0.16	0.64
Ours	**0.51**	**1.82**	**1.35**	**15.99**	0.39	**0.77**

Table 8: Comparison against NEON ALS canopy height in the United States for 2020, using MAE (m), MSE (m^2), RMSE (m), MAPE (%), and R^2. The epoch-corrected Potapov row adds the *ECHOSAT* 2019–2020 increment to the 2019 Potapov map.

Method	MAE ↓	MSE ↓	RMSE ↓	MAPE ↓	R^2 ↑
Potapov et al. (2021)	5.299	48.809	6.986	38.557	−0.352
Potapov et al. (2021), epoch-corrected	5.286	48.593	6.971	38.497	−0.346
Lang et al. (2023)	4.426	31.074	5.574	37.742	0.137
Pauls et al. (2024)	3.649	22.477	4.741	24.841	0.377
Tolan et al. (2024)	6.331	58.803	7.668	43.626	−0.629
Our	**2.949**	**15.081**	**3.883**	**20.713**	**0.582**

A.2.6 Comparison to NEON ALS

Validating large-scale tree-height products remains challenging. GEDI provides global coverage, though with some noise and uncertainty, while airborne laser scanning (ALS) offers higher accuracy but is spatially limited. We therefore repeat the 2020 comparison of Table 1 on independent ALS canopy heights from the National Ecological Observatory Network (NEON) across sites in the United States (National Ecological Observatory Network (NEON), 2026). Table 8 reports the metrics. The ranking from Table 1 is preserved: *ECHOSAT* remains the most accurate map, followed by Pauls et al. (2024), Lang et al. (2023), Potapov et al. (2021), and Tolan et al. (2024). Figure 12 shows the corresponding density scatterplots against NEON ALS p98, the 98th percentile canopy height. They closely resemble the GEDI scatterplots in Figure 6, with substantially less noise about both axes and the diagonal. We take this as evidence that the improvement is not only a better fit to GEDI-specific noise.

The Potapov map is for 2019, whereas the NEON comparison is for 2020. We also report an epoch-corrected version in which the *ECHOSAT* increment (*ECHOSAT* 2020 − *ECHOSAT* 2019) is added to the Potapov map, so that this baseline is not penalized for the one-year offset. The correction changes the result only marginally (MAE $5.299\,\mathrm{m}$ to $5.286\,\mathrm{m}$) and does not affect the ranking. This is consistent with the small year-to-year increments in Section 4.2: one-year *ECHOSAT* increments are too small to move a baseline far, while the direction of the correction is the expected one.

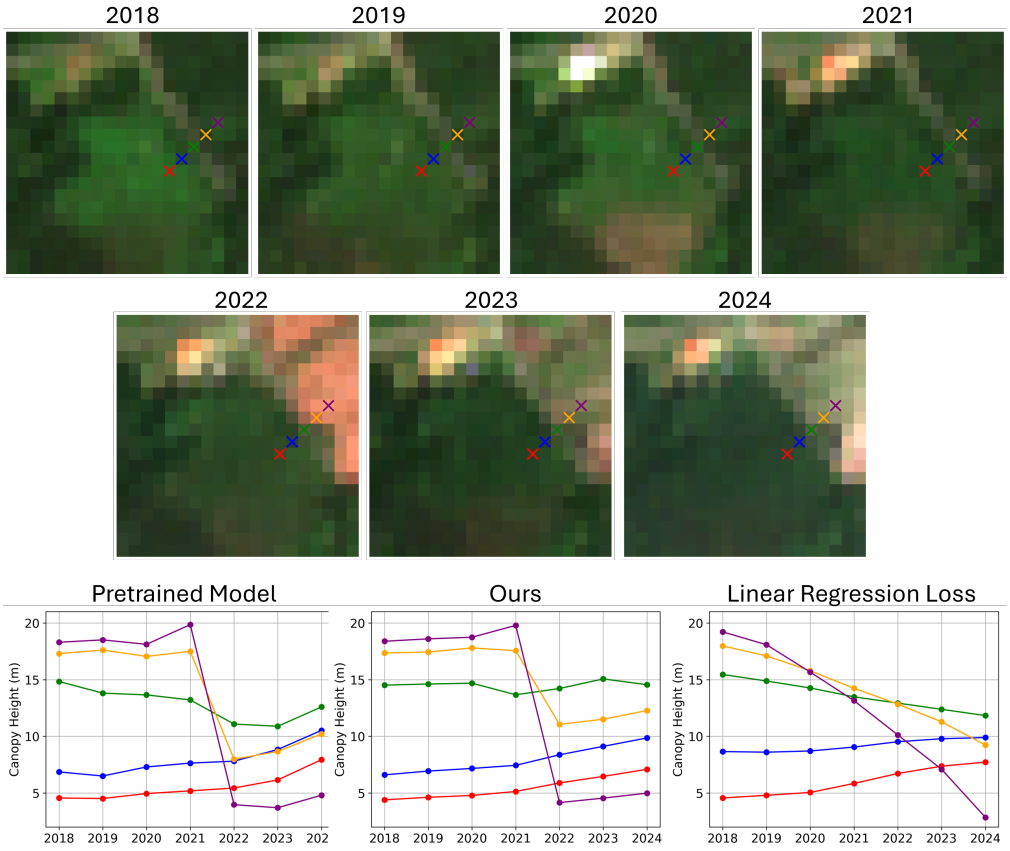


Figure 9: Sentinel-2 time series and temporal tree height predictions for five neighbouring pixels around a disturbance for the pretrained model (left), our model (middle) and the model fine-tuned with simple linear regression model (right). The simple linear regression model is fine-tuned on pseudo-labels produced by a linear interpolation of the pretrained model's predictions.

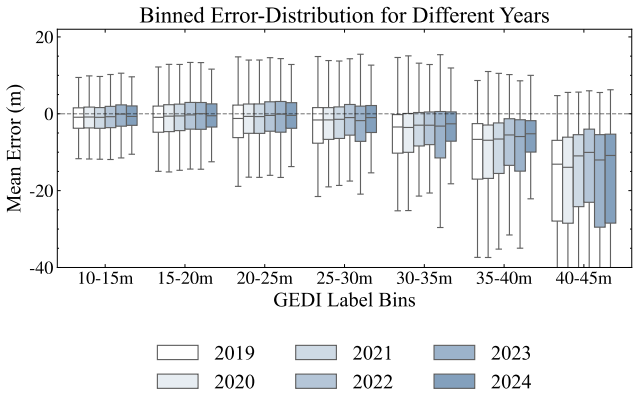


Figure 10: Error distribution analysis across height classes (5 m bins) for *ECHOSAT* predictions and different years. Boxplots show mean absolute error for each height class. The errors for smaller trees are very similar, with higher trees showing earlier years as being slightly more difficult than recent years.

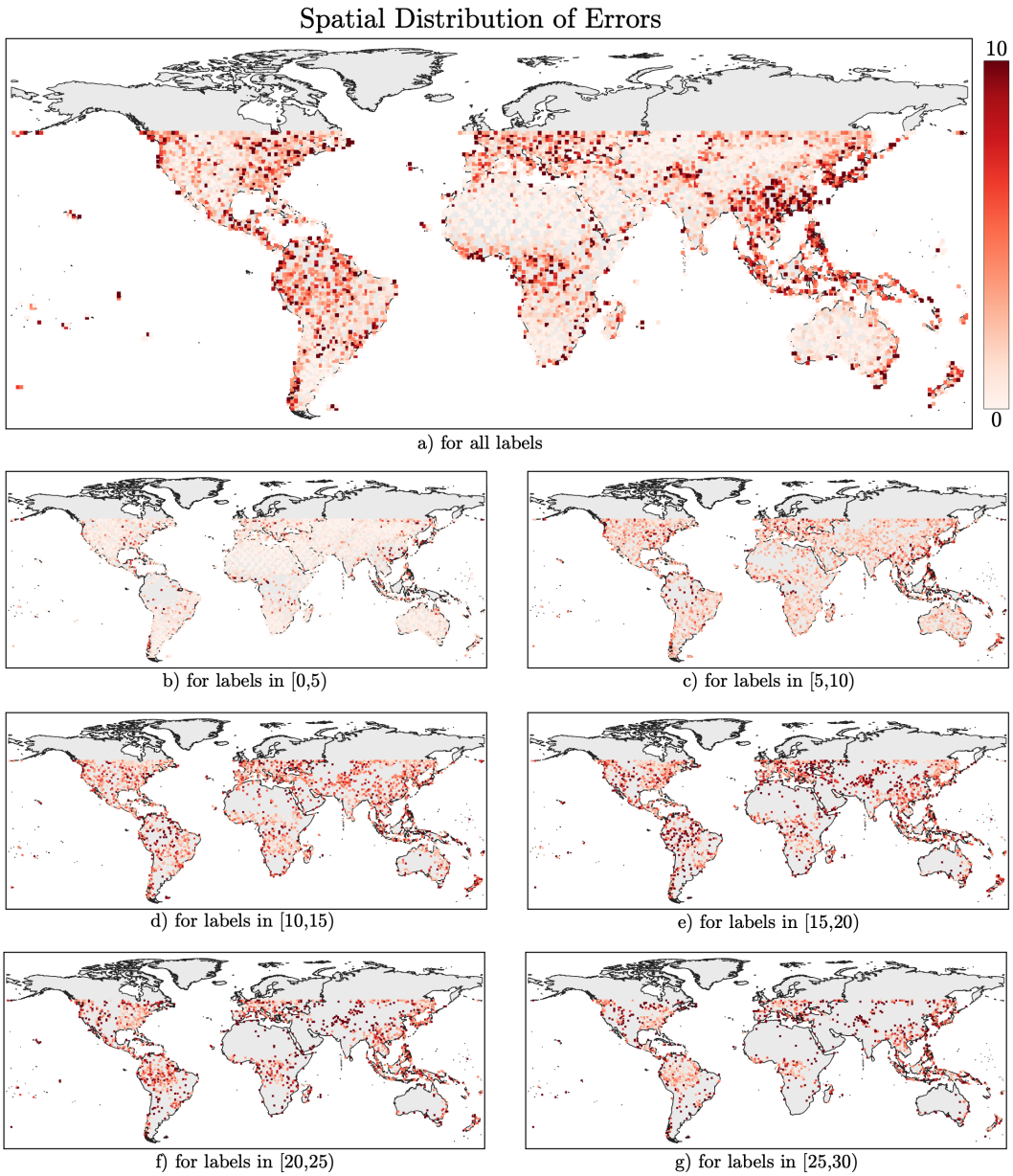
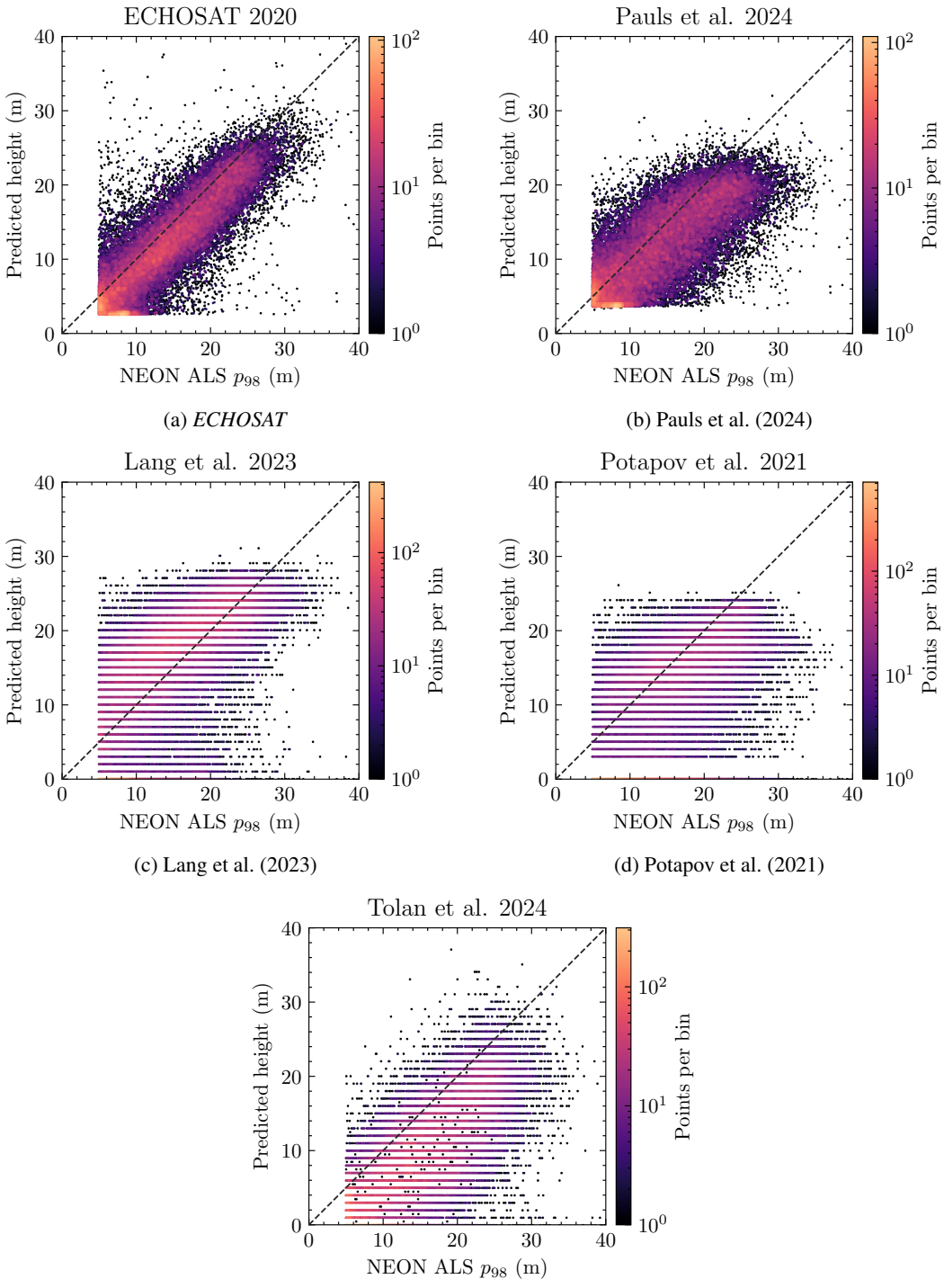


Figure 11: Spatial distribution of errors for all labels (a) and different height bins (b-g). a) strongly correlates with known areas of taller trees (i.e. higher errors, see Figure 6), small trees (b-c) generally have low errors globally, whereas errors systematically grow with tree height (d-g). Interestingly, (1) the east coast of North America has lower errors compared to other areas and (2) the Amazonas rainforest and Congo Basin, known for tall trees, have smaller errors (g) compared to other regions.



(a) *ECHOSAT*

(b) Pauls et al. (2024)

(c) Lang et al. (2023)

(d) Potapov et al. (2021)

(e) Tolan et al. (2024)

Figure 12: Density scatterplots of 2020 predicted height against NEON ALS p98 in the United States. Color indicates the number of points per bin. The *ECHOSAT* panel is the narrowest about the diagonal, in line with Table 8.

A.3 MODEL HYPERPARAMETERS

Table 9: Training hyperparameters. Differences between pretraining and fine-tuning are highlighted.

Parameter	**Symbol**	**Value**	
		Pretraining	*Finetuning*
Number of years	Y	7	
Number of timesteps	T	84 ($= Y \cdot Months$)	
Number of input channels	C	18	
Embedding dimension	E	72	
Height / Width (in pixels)	H / W	96	
Encoder depths	depth$[l_{\text{enc}}]$	$[6, 4, 4, 6]$	
Decoder depths	depth$[l_{\text{dec}}]$	$[4, 6, 8, 16]$	
Attention heads		$[4, 8, 12, 24]$	
Temporal window size		2	
Spatial window size		6	
Embedding patch size	$P_H \times P_W \times P_T$	$1 \times 1 \times 1$	
Optimizer		`AdamW`	
Maximum learning rate		1×10^{-4}	3×10^{-3}
Learning rate linear warmup		30%	
Learning rate schedule		Cosine Annealing	
Gradient clipping		1	
Number of iterations		539k	179k
Loss		$\mathcal{L}_{\text{huber}}(\cdot)$	$\mathcal{L}_{\text{growth}}(\cdot)$
Batch size		16	8