Canopy Tree Height Estimation Using Quantile Regression: Modeling and Evaluating Uncertainty in Remote Sensing

Karsten Schrödter
University of Münster

Jan Pauls
University of Münster

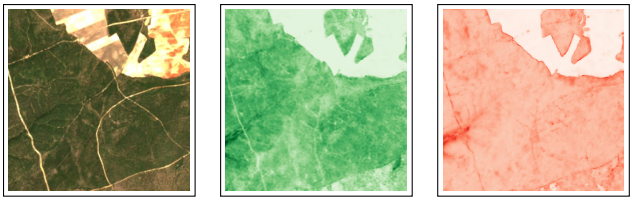
Fabian Gieseke
University of Münster
University of Copenhagen

Abstract

Accurate tree height estimation is vital for ecological monitoring and biomass assessment. We apply quantile regression to existing tree height estimation models based on satellite data to incorporate uncertainty quantification. Most current approaches for tree height estimation rely on point predictions, which limits their applicability in risk-sensitive scenarios. In this work, we show that, with minor modifications of a given prediction head, existing models can be adapted to provide statistically calibrated uncertainty estimates via quantile regression. Furthermore, we demonstrate how our results correlate with known challenges in remote sensing (e.g., terrain complexity, vegetation heterogeneity), indicating that the model is less confident in more challenging conditions.

1 Introduction

Monitoring forests is an important component of assessing the effort required to comply with the Paris Climate Agreement.[1] In recent years, the availability of satellite remote sensing data at high resolution and at a global scale as well as advances in machine learning have paved the way for an automated forest monitoring. A typical building block for analyses in this domain is the estimation of (canopy) tree height from remote sensing data, which then serves as a proxy for follow-up analyses such as so-called above-ground biomass estimation (Schwartz et al., 2023) and, thus, quantification of stored carbon.



(a) (b) (c)

Figure 1: Tree height estimation: An input image (a) can be used to produce a tree height map (b) using a tree height model (intensity of green increases with higher tree height estimation). Most studies focus on point estimates. In Figure (c), associated uncertainty estimates are provided (intensity of red increases with higher uncertainty). It can be seen that the model is less confident at forest borders and along cutoff stripes.

An example of such a tree height estimation scenario is shown in Figure 1. Here, the input satellite data (a) are used to obtain an estimate (b) for the tree height in the corresponding area. Most recent studies have focused on using point-estimators to produce canopy height maps from country-to-global scale (Schwartz et al., 2022; Tolan et al., 2024; Pauls et al., 2024, 2025; Potapov et al., 2021). However, valid uncertainty estimation for canopy height maps is rare, with Lang et al. (2023) among the few exceptions. Figure 1 (c) shows an uncertainty map for the height estimations given in Figure (b). It can be seen, for instance, that the model is less confident at forest borders. Uncertainty quantification is essential for decision- and policy-making, where worst-case scenarios must be incorporated into climate simulations and general risk assessments (Smith and Stern, 2011). In forest monitoring, quantifying uncertainty in canopy-height estimates enables its propagation through non-linear downstream models of above-ground biomass and carbon storage, underpinning robust derived products such as CO_2 certificates.

[1] For instance, on a global scale, forests absorb 3.5 ± 0.4 Pg of carbon per year (Pan et al., 2024).

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

In this work, we use quantile regression to incorporate uncertainty quantification into existing pretrained tree height models, evaluate their effectiveness using recently proposed architectures, and analyze the resulting uncertainty maps on real-world data. On held-out validation sets, we show that the predicted uncertainties are well calibrated and we characterize recurrent conditions that systematically lead to elevated uncertainty. We expect our findings to advance the use of uncertainty quantification for this task, and to pave the way for a more reliable use of height maps for various downstream tasks, including above-ground biomass estimation and carbon accounting.

2 Background

We start by providing details about canopy tree height estimation along with uncertainty quantification for deep learning models.

2.1 Canopy Height Estimation

Above-ground biomass is essential for carbon storage estimation and forest monitoring. However, not much ground-truth data is directly available, so tree height as an approximator is frequently used (Schwartz et al., 2023). Early approaches relied on manual field sampling and spatial interpolation, which were both costly and limited in coverage. The advent of free and open satellite imagery from missions such as Landsat and Sentinel-2 enabled tree height estimation at high resolution and continental to global scales.

Modern canopy height estimation leverages machine learning models trained on satellite imagery, such as the optical images from Sentinel-2, with ground truth labels derived from spaceborne LiDAR missions, like the Global Ecosystem Dynamics Investigation (GEDI) (Dubayah et al., 2022). GEDI and ICESat-2 (Neumann et al., 2019) missions provide sparse but accurate tree height measurements on a global scale. GEDI, a full-waveform LiDAR instrument aboard the International Station (ISS), records footprints with a beam diameter of approximately 25 m, spaced 60 m apart along-track, across 8 parallel tracks. Geolocation is determined via GPS, star trackers, and instrument orientation data, achieving a reported 1-sigma accuracy of approximately 10.2 m, though accuracy varies across individual tracks.

The spatial resolution of canopy height products has steadily improved alongside advances in satellite imagery availability. Early studies build on 30 m resolution data from Landsat (available for 30+ years). More recent work exploits 10 m resolution Sentinel-2 imagery (openly available since 2015), while commercial satellites provide 3 m daily coverage and sub-meter resolution for targeted regions. Current studies span multiple spatial scales: regional (Rolf et al., 2024), national (Schwartz et al., 2022; Su et al., 2025), continental (Pauls et al., 2025; Liu et al., 2023), and global (Tolan et al., 2024; Lang et al., 2023; Pauls et al., 2024; Potapov et al., 2021), with resolutions ranging from 30 m to sub-meter scales.

Despite continuous improvements in point prediction accuracy, uncertainty quantification remains largely unaddressed. To our knowledge, Lang et al. (2023) is the only large-scale study that provides uncertainty estimates. This represents a critical gap, as uncertainty estimates are essential for downstream applications: they enable reliable risk assessment in carbon offset programs, inform targeted allocation of resources for field campaigns, and allow propagation of prediction uncertainty through non-linear biomass and carbon storage models (Yambayamba et al., 2025; Lin et al., 2023; Bian and Xia, 2023; Candelas Bielza et al., 2025). Without calibrated uncertainty estimates, canopy height predictions cannot be safely integrated into climate policy decisions that require worst-case scenario analysis (Besic et al., 2025).

2.2 Uncertainty in Deep Learning

Methods for uncertainty quantification can roughly be put into 4 categories, see Gawlikowski et al. (2023): Bayesian approaches, ensemble methods, test-time augmentation and single deterministic methods. Bayesian models put distributions on model parameters, leading to posterior-predictive distributions. Ensembles aim to infer uncertainty having multiple models that are initialized independently. Bayesian and ensemble approaches are highly compute-intense, that is why we do not apply them for canopy height estimation. Test-time augmentation methods apply multiple data augmentation techniques to input data at prediction time, often used in healthcare applications. This is challenging in the context of canopy height estimation, as there are few well established augmentations that keep canopy height labels unchanged. Single deterministic methods summarize all methods, where the uncertainty is generated in a single forward pass of a deterministic network. An example is Gaussian regression, where the output is modeled to follow a Gaussian distribution by estimating mean and variance and is trained to minimize the negative log-likelihood, a method that was introduced latest in 1994, see Nix and Weigend (1994).

By now, Lang et al. (2023) proposed the only approach for incorporating uncertainty quantification into a canopy height model by using an ensemble of Gaussian regression models. More single determinis-

tic methods are tested in other biological applications of image-to-image regression tasks. Regression on the magnitude of residuals aims to predict the model's error for each prediction. Moreover, regression can be re-framed as a classification task and a softmax distribution can be used for uncertainty quantification. Another well-known technique is quantile regression, where an asymmetric loss function is used to predict conditional quantiles. All these methods are tested in Angelopoulos et al. (2022) for different image-to-image regression tasks, with the finding that quantile regression performs best among these methods. There are theoretical results that guarantee good behaviour of quantile regression under mild assumptions, see Steinwart and Christmann (2011). Quantile regression is reported to show good results on image-to-image tasks, so we decided to apply it to canopy height estimation.

3 Approach

Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and output space of our supervised learning task. An input datum consists of T images of height H and width W, each having C channels. Therefore, the input space will be denoted as $\mathcal{X} = \mathbb{R}^{T\times C\times H\times W}$. An output consists of N images of the same size, while a label consists of a single image of the same size. Thus, the label space will be denoted by $\mathcal{Y} = \mathbb{R}^{H\times W}$. Each output image aims to train a quantile canopy height model $f : \mathcal{X} \to \mathcal{Y}$ by minimizing some loss on training data $\mathcal{D} = ((X_1, Y_1), \ldots, (X_M, Y_M)) \in (\mathcal{X} \times \mathcal{Y})^M$. Note that since the labels in $\mathcal{D}$ are generated from GEDI measurements, it contains sparse information, see the right part of Figure 6. Having no information on a label $Y \in \mathcal{Y}$ at an index (h, w) is encoded by $Y_{h,w} = 0$. For a label $Y \in \mathcal{Y}$ we define its sparse notation by

$$\mathbb{Y} := \{(h, w, y) \mid 1 \le h \le H, 1 \le w \le W, Y_{h,w} = y > 0\},$$

where $(h, w, y) \in \mathbb{Y}$ reads as the ground truth at pixel (h, w) is y. We denote the space of sparse labels by $\mathcal{Y}_{\text{sparse}}$. The sparse notation is introduced to facilitate the notation of the loss functions in our context.

3.1 Pinball Loss

The basic idea behind quantile regression is to use the pinball loss, which penalizes over- and underprediction asymmetrically. Let $\tau \in (0, 1)$ be a target quantile then the pinball loss $\mathcal{L}_\tau : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+ := [0, \infty)$ is defined by

$$\mathcal{L}_\tau(y, \hat{y}) := \begin{cases} \tau \cdot (y - \hat{y}) & \text{if } \hat{y} \le y \\ (1-\tau) \cdot (\hat{y} - y) & \text{if } \hat{y} > y \end{cases}, \quad (1)$$

where $y \in \mathbb{R}$ is the ground truth value and $\hat{y} \in \mathbb{R}$ the prediction. For example, if $\tau = 0.2$, note that underprediction $(\hat{y} \le y)$ is penalized with a small factor of $\tau = 0.2$, while overshooting $(\hat{y} > y)$ is penalized with a higher factor of $1 - \tau = 0.8$.

Let P be a distribution on $\mathcal{X} \times \mathbb{R}$ and define the conditional τ-quantile of $x \in \mathcal{X}$ as

$$Q_\tau(x) := \inf\{q \in \mathbb{R} \mid P(y \le q \mid x) \ge \tau\}.$$

Then it is well-known, see e.g. Steinwart and Christmann (2011), that the function Q_τ sending $x \in \mathcal{X}$ to $Q_\tau(x)$ is the minimizer among all measurable functions $f : \mathcal{X} \to \mathbb{R}$ for the risk of the pinball loss, i.e.

$$Q_\tau = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}\times\mathbb{R}} \mathcal{L}_\tau(y, f(x)) dP(x, y).$$

3.2 Pixelwise Quantile Regression

In the context of image-to-image regression, the loss for pixelwise quantile regression averages the pinball loss at every pixel, as in Angelopoulos et al. (2022). In our context, we only have sparse labels and thus only average over the pixels with label. More precisely, let $Y \in \mathbb{R}^{H\times W}$ be a prediction and denote the ground truth value in sparse notation by $\mathbb{Y} \in \mathcal{Y}_{\text{sparse}}$. Then the sparse pixelwise pinball loss is defined as

$$\mathcal{L}_\tau(\mathbb{Y}, Y) := \frac{1}{\#\mathbb{Y}} \sum_{(h,w,y)\in\mathbb{Y}} \mathcal{L}_\tau(y, Y_{h,w}) \in \mathbb{R}^+. \quad (2)$$

With our model, we train multiple output images simultaneously on various quantiles $\tau_1, \ldots, \tau_N \in (0, 1)$. The final loss function is the average of all pinball losses of the output images, more precisely, denote the n-th output image of the model by $Y^{(n)} \in \mathbb{R}^{H\times W}$ and consider a sparse ground truth label $\mathbb{Y} \in \mathcal{Y}_{\text{sparse}}$. Then the loss function $\mathcal{L}_{\bar{\tau}}$ is

$$\mathcal{L}_{\bar{\tau}}(\mathbb{Y}, \mathbf{Y}) := \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{\tau_n}(\mathbb{Y}, Y^{(n)}) \in \mathbb{R}^+, \quad (3)$$

where $\bar{\tau} = (\tau_1, \ldots, \tau_N) \in (0, 1)^N$ is the vector of quantiles and $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(N)}) \in \mathbb{R}^{H\times W\times N}$ the concatenated output.

3.3 Shift-Resilient Loss

As in Pauls et al. (2024), we use a shifted variant of our loss. GEDI labels can be divided into tracks, i.e. into measurements that are taken on the same overflight path. The basic observation for the shift-resilient loss is that geolocation errors tend to be constant along tracks. Therefore, the idea is to allow the model to shift the labels within a track by a small offset, if this decrease the overall loss of the track.

More precisely, a sparse label $\mathbb{Y} \in \mathcal{Y}_{\text{sparse}}$ can be partitioned into GEDI tracks that can be describe by $t_1, \ldots, t_I \in \mathcal{Y}_{\text{sparse}}$ such that

$$t_i \subset \mathbb{Y},\ \bigcup_{i=1}^{I} t_i = \mathbb{Y},\ t_i \cap t_j = \emptyset$$

for $1 \leq i \neq j \leq I$. All points of the same track are approximately on one line. For a track $t \subset \mathbb{Y}$ and a shifting direction $\delta = (\delta_h, \delta_w) \in \mathbb{Z}^2$ we define the shifted track $t(\delta)$ by

$$t(\delta) \coloneqq \{(h + \delta_h, w + \delta_w, y) \mid (h, w, y) \in t\}.$$

The shifted (sh) pixelwise quantile regression loss for a track t, a quantile vector $\bar{\tau} = (\tau_1, \ldots, \tau_N) \in (0,1)^N$ and a multi-image output $\mathbf{Y} \in \mathbb{R}^{H \times W \times N}$ is given by

$$\mathcal{L}_{\bar{\tau}}^{\text{sh}}(t, \mathbf{Y}) \coloneqq \min_{\delta \in \{-1,0,1\}^2} \mathcal{L}_{\bar{\tau}}(t(\delta), \mathbf{Y}). \tag{4}$$

In particular, the track is allowed to be shifted by at most one pixel in each direction to find the best fitting predictions. The shifted pixelwise regression loss for a sparse label $\mathbb{Y}$ containing of tracks $t_1, \ldots, t_I \in \mathcal{Y}_{\text{sparse}}$ is given by the average of all track losses, i.e.

$$\mathcal{L}_{\bar{\tau}}^{\text{sh}}(\mathbb{Y}, \mathbf{Y}) \coloneqq \frac{1}{I} \sum_{i=1}^{I} \mathcal{L}_{\bar{\tau}}^{\text{sh}}(t_i, \mathbf{Y}). \tag{5}$$

4 Experiments

We incorporate quantile regression into an existing tree canopy height model by adding an uncertainty head to the model. After fine-tuning of the model, we evaluate the point-estimator and the uncertainty estimates using metrics and qualitative comparisons to the model from Lang et al. (2023).

4.1 Experimental Setup

To evaluate our approach, we perform several experiments testing the applicability to tree canopy height prediction on continental-scale. We use the model from Pauls et al. (2025) and use an additional uncertainty head, which is trained using shifted pixelwise pinball loss. We fine-tune both the original and the uncertainty head on the same dataset for a quarter of the pretraining epochs and evaluate the model on a test dataset with 250 samples from 2020, containing approximately 49k GEDI labels. We compare our results to Lang et al. (2023) and show analysis on uncertainty effects near forest borders and mountainous terrain. Additionally, we performed an ablation study on exchanging quantile regression by either Gaussian or Log-Gaussian regression, which aim to maximize the average likelihood of uncertainty predictions using a Gaussian, resp. Log-Gaussian, distribution. Gaussian regression is also used in Lang et al. (2023) for every ensemble member.

In the following, we shortly describe the model, our fine-tuning strategy and the evaluation metrics.

4.1.1 Data and Architecture

Pauls et al. (2025) propose to use a Unet model that is adopted for three-dimensional input, i.e. processing not just one satellite image, but rather a time-series. More precisely, they take a time-series of $T = 12$ monthly Sentinel-2 images of size $H = W = 256$ and concatenate a yearly composite of Sentinel-1 radar data channel-wise, resulting in $C = 16$ channels. Their ground-truth data stems from GEDI measurements that are filtered using several quality criteria to limit the influence from noisy labels.

We use the pretrained head and fine-tune it to predict a 50%-quantile, usable as point-estimator. Additionally, we aim to train the model to predict 10 more quantiles, namely:

$$\bar{\tau} = (0.05, 0.1, 0.15, 0.2, 0.25, 0.75, 0.8, 0.85, 0.9, 0.95)$$

In total, the model predicts $N = 11$ channels. We adapt the architecture from Pauls et al. (2025) by adding a new head, see Appendix A for details. In the original architecture the head consisted of one 1×1 convolution that reduces the channels from 64 to 1. We build a second path and append an uncertainty head to the same 64 input channels. The head consists of two convolution blocks followed by a similar 1×1 convolution to reduce the channels from 64 to 10. One convolution block consists of a 2D convolution with a kernel of size 3×3 and same padding that do not alter the number of channels.

4.1.2 Training Setting

The model from Pauls et al. (2025) was pretrained on Huber loss using the Adam optimizer (Kingma, 2014) with an initial learning rate 0.001, weight decay of 0.01 and gradient clipping at 1.0. Training was done for a total of 400,000 iterations with batch size 16 and a linear learning rate scheduler with 10% warmup iterations was used. Our fine-tuning follows a similar procedure, however the shifted pixelwise pinball loss is used and we only fine-tune for 2 epochs using a batch size of 5 and freeze everything except the two heads (point-estimator and uncertainty head).

4.1.3 Uncertainty Quantification Evaluation

We evaluate our predictions using several metrics that measure prediction intervals width, coverage and

more. Here we introduce these concepts:

Let $\mathcal{T} = ((X_1, \mathbb{Y}_1), \ldots, (X_M, \mathbb{Y}_M)) \in (\mathcal{X} \times \mathcal{Y})^M$ be a test dataset, $\tau \in (0,1)$ be the desired quantile and $f_\tau : \mathcal{X} \to \mathcal{Y}$ be a function – e.g. a trained neural network – to predict the τ-conditional quantile. We evaluate the predictions for a quantile $\tau \in (0,1)$ by calculating the Empirical Coverage (EC), defined by

$$\mathrm{EC}_\tau = \frac{\sum_{m=1}^{M} \#\{(h,w,y) \in \mathbb{Y}_m \mid f_\tau(X_m)_{h,w} \geq y\}}{\sum_{m=1}^{M} \#\mathbb{Y}_m}, \quad (6)$$

which is essentially the proportion of labels that are overestimated by f_τ among all labeled points.

Furthermore, we form prediction intervals (PI) for uncertainty level $\alpha \in (0,1)$ by predicting lower and upper quantiles τ_l and τ_h, defined by

$$\tau_l \coloneqq 0.5 - \alpha/2 \text{ and } \tau_h \coloneqq 0.5 + \alpha/2,$$

and span an interval around the corresponding predictions, i.e.

$$PI_\alpha(X) = [f_{\tau_l}(X), f_{\tau_h}(X)]$$

for an input image $X \in \mathcal{X}$. Furthermore, we set the Prediction Interval Width (PIW) to

$$\mathrm{PIW}_\alpha(X) \coloneqq f_{\tau_h}(X) - f_{\tau_l}(X) \in \mathbb{R}^{H \times W} \quad (7)$$

and use established uncertainty metrics such as Mean Prediction Interval Width (MPIW) and Prediction Interval Coverage Probability (PICP), defined as

$$\mathrm{MPIW}_\alpha \coloneqq \frac{\sum_{m=1}^{M} \sum_{(h,w,y) \in \mathbb{Y}_m} \mathrm{PIW}_\alpha(X_m)_{h,w}}{\sum_{m=1}^{M} \#\mathbb{Y}_m}, \quad (8)$$

$$\mathrm{PICP}_\alpha \coloneqq \frac{\sum_{m=1}^{M} \#\{(h,w,y) \in \mathbb{Y}_m \mid y \in \mathrm{PI}_\alpha(X_m)\}}{\sum_{m=1}^{M} \#\mathbb{Y}_m}. \quad (9)$$

4.2 Results

As our approach serves as a mere add-on, we want to ensure that the model's point-prediction is not worsened after our architectural changes and fine-tuning. Section 4.2.1 shows experiments comparing both stages and Lang et al.'s prediction to the available ground-truth measurements. Sections 4.2.2 - 4.2.5 analyze the Prediction Interval Width from different perspectives, Sections 4.2.6 and 4.2.7 examine the uncertainty at forest borders and mountainous areas, and Section 4.2.8 compares the visual prediction quality. Appendix B presents the results of the ablation study comparing quantile regression with Gaussian and Log-Gaussian regression. Quantile regression outperformed both alternatives in terms of point estimation and uncertainty quantification. We therefore focus on quantile regression in the following.

Furthermore, we refer to Appendix C for an analysis of the computational overhead of our architecture compared to the model from Pauls et al. (2025).

4.2.1 Point Estimation Comparison

The point-estimator results of the three models (Lang et al. (2023), Pauls et al. (2025), and Ours) are summarized in Table 1. The results are in line with findings in Pauls et al. (2025) that the model from Pauls et al. yields better metrics than Lang et al.. Note that fine-tuning the prediction head using the pinball loss for $\tau = 0.5$, i.e. essentially using Mean Absolute Error loss, did not change metrics drastically compared to Pauls et al.. However, it is worth mentioning that the prediction of Pauls et al.'s model, which is trained using the Huber loss, is above the ground truth GEDI value in about 58% of the cases, while the fine-tuned 50%-quantile prediction is well-calibrated. Note that the EC values for Pauls et al. and Lang et al. are only given as a reference, since both models are not trained to predict a 50%-quantile, but rather a mean in the case of Lang et al., or a combination of mean and median (Pauls et al.).

Table 1: Results on point-estimators, measured using Mean Squared Error (MSE, m^2), Mean Absolute Error (MAE, m), R^2 and Empirical Coverage (EC).

Model	MSE ↓	MAE ↓	R^2 ↑	EC
Lang et al.	38.32	4.25	0.55	0.47
Pauls et al.	20.64	1.90	0.76	0.58
Ours	20.81	1.90	0.76	0.50

4.2.2 Uncertainty Quantification Evaluation

Comparing the EC across all trained quantile-levels, depicted in Figure 2, reveals that Lang et al. (2023) slightly underestimates the smaller quantiles, where our model's quantile predictions are consistently a bit too high. Noticeably, on higher quantiles, our model achieves similar performance, but underestimates the quantiles, where Lang et al. (2023) is covering less data, resulting in the 90%-quantile output only overshooting 67% of the labels. Table 2 shows MPIW and

Table 2: Comparison of Mean Prediction Interval Width (MPIW) and Prediction Interval Coverage Probability (PICP) for multiple uncertainty levels α for Lang et al.'s model and our model.

Metric	MPIW					PICP				
α	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
Lang et. al.	5.54	6.87	8.39	10.22	12.76	0.37	0.45	0.52	0.58	0.64
Ours	2.48	3.07	3.79	4.77	6.47	0.48	0.58	0.67	0.78	0.88

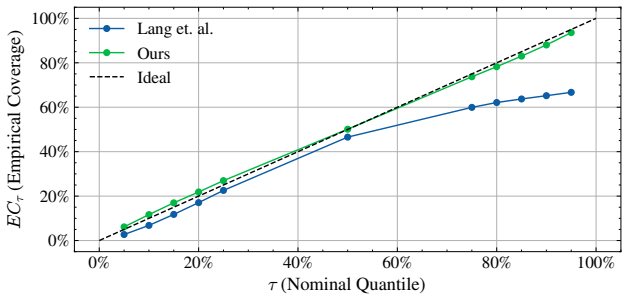


Figure 2: Empirical Coverage vs Nominal Quantile for Lang et al.'s and our model.

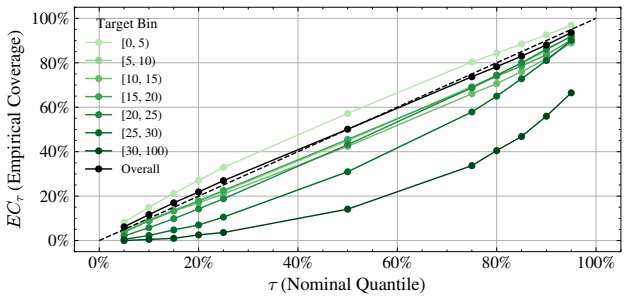


Figure 3: Empirical Coverage vs Nominal Quantile for multiple target bins for our model.

PICP across different uncertainty levels α. Our model consistently achieves better PICP values with prediction intervals roughly half as wide as those of Lang et al., whose prediction intervals moreover cover noticeably fewer labels than expected.

4.2.3 Empirical Coverage across Target Bins

We evaluate the empirical coverage of predicted quantiles from our models across different target bins, results are depicted in Figure 3. For the 50%-quantile prediction, i.e. our point-estimator, we see the known pattern that tall trees are underpredicted, see also the scatter plot in Figure 5. Additionally, here we see that all quantiles tend to overestimate small vegetation below 5 m and underestimate taller trees. Note that for trees below 25 m, the empirical coverage lies within a reasonable range of less than 5% away from the expected value. For trees taller than 30 m, even the 95%-quantile prediction covers less than 70% of the trees. Note that correcting for these miscalibrated empirical coverages is non-trivial, as the miscalibration only manifests when predictions are grouped by target bins. When grouped by prediction bins instead, all groups are well-calibrated, compare Appendix D, Figure 11.

4.2.4 Prediction Interval Asymmetry

To test for asymmetry of the prediction intervals around the median prediction, Figure 4 shows boxplots of the distances from the lower and upper edges of the prediction interval to the median prediction.

First, the distances increase with the confidence level α, as expected, since higher-coverage prediction intervals are wider. Second, across all confidence levels, the upper edge is consistently further from the median than the lower edge, indicating a right-skewed predictive distribution. This asymmetry is particularly pronounced for the 90% prediction interval, where the median distance of the upper edge ($\sim$2.54 m) is roughly three times that of the lower edge ($\sim$0.8 m). Notably, models that assume a symmetric distribution, such as the Gaussian regression approach of Lang et al. (2023), are inherently unable to capture this behavior.

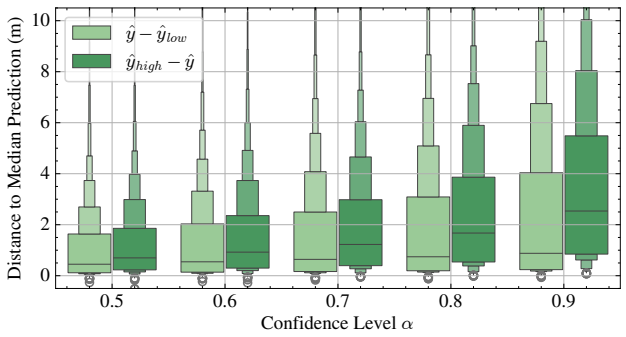


Figure 4: Boxplot of distance of quantile predictions to point-estimate, i.e. median prediction. Per confidence level $\alpha \in [0.5, 0.6, 0.7, 0.8, 0.9]$, the prediction interval at level α is written as $PIW_\alpha = [\hat{y}_{low}, \hat{y}_{high}] \subset \mathbb{R}$, while the median prediction is denoted $\hat{y} \in \mathbb{R}$.

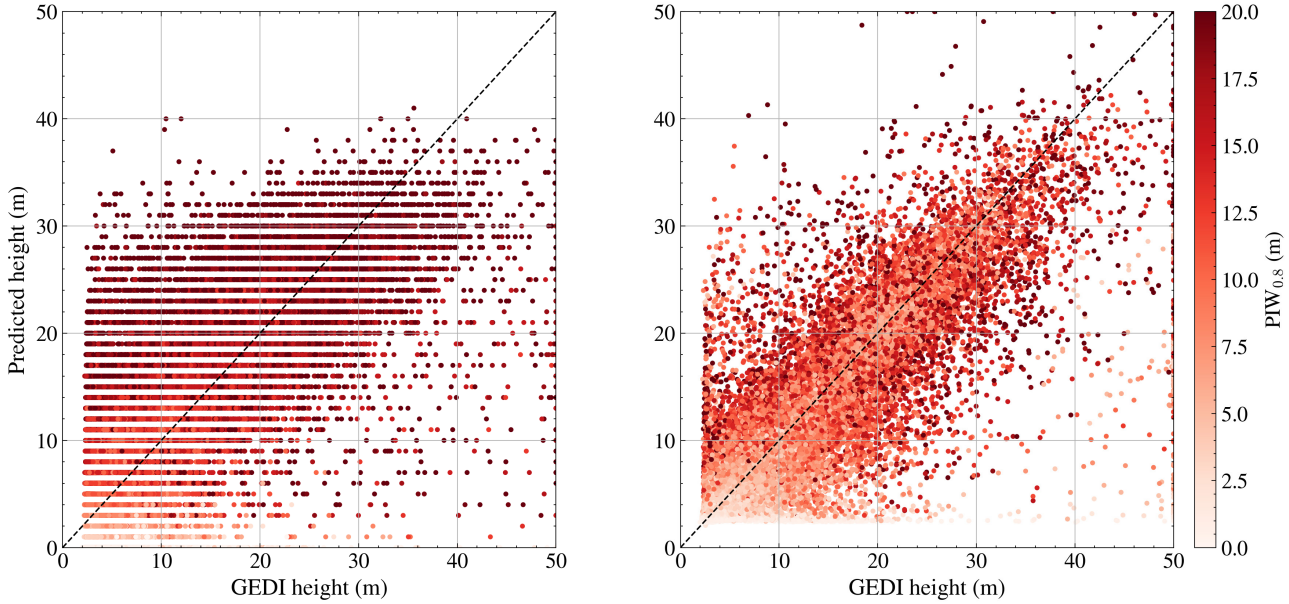


Figure 5: Scatter plot of GEDI height against predicted height with color dependent on the prediction interval width at level $\alpha = 0.8$. All heights are clipped into $[0\,\text{m}, 50\,\text{m}]$. Left: Lang et al.'s model. Right: Our model.

4.2.5 Correlation of Prediction and Uncertainty

Figure 5 compares GEDI reference height against predicted height with color-marked uncertainty, where a darker red indicates lower confidence. Note that Lang et al.'s prediction are available in whole meters, leading to the horizontal stripes in the Figure. Besides the observation that predictions of our model are more accurate overall, it is noticeable that for Lang et al.'s model a small predicted uncertainty occurs only when the prediction is also low. In particular, the correlation between height prediction and uncertainty estimation is higher for Lang et al. (Pearson correlation 0.85 vs 0.72 for our model), showing that uncertainty tends to rise with higher predictions. For our model on contrast, there is a clear trend visible that points further away from the diagonal, i.e. points with higher errors, also have higher uncertainty, which is in line with what we expect from a good uncertainty model. However, there are also counterexamples. Note that there are some points, where the model is highly confident about the prediction of less than 10 m, but the corresponding GEDI label is above 30 m. An investigation of such situations indicate that these points might suffer from label noise, see Figure 6 for an example. The marked points lie on cropland, with a low prediction below 10 m of the 90%-conditional quantile, but have GEDI labels higher than 30 m. Therefore, wrong predictions with high confidence might be an indicator of false labels. An approach to use this to detect and remove false labels from the dataset during training remains a subject of future work. Interestingly, there are much fewer points with a small GEDI label and a high prediction that the model is confident about.

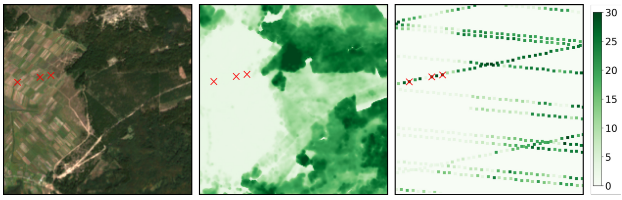


Figure 6: Noisy GEDI labels marked in red: 90%-quantile prediction below 10 m and GEDI labels above 30 m. Left: Satellite image, Middle: 90%-quantile canopy height prediction, Right: GEDI labels

4.2.6 Higher Uncertainty at Forest Borders

In this section, we investigate the predicted uncertainty at forest borders. We define a pixel to lie at the border of a forest if the minimal and maximal predictions in the 3×3 surrounding of the pixel differ by more than 10 m. Figure 7 shows a boxplot of PIW between forest border pixels and non-forest border pixels. It is clearly visible that there is a huge difference in the PIW between pixels at the forest border and non-forest border pixels, meaning that, overall, our model is much less confident around forest borders. Additionally, the PICP for uncertainty level $\alpha = 0.8$ is about 78% for non-forest border pixels and only 70% for forest border pixels, indicating that uncertainty estimation is more complicated at forest borders. We hypothesize that this is at least partly a consequence of the geolocation error of GEDI measurements. Note that a geoloca-

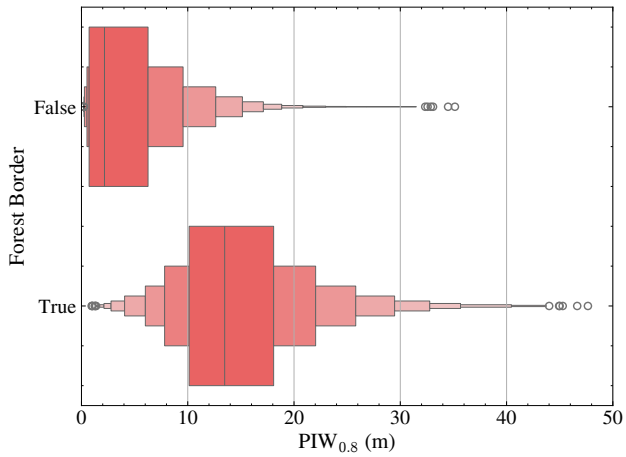


Figure 7: Boxplot of Prediction Interval Width at level $\alpha = 0.8$ in dependence of being at a forest border. Note that less than 5% of the pixels are located at forest borders.

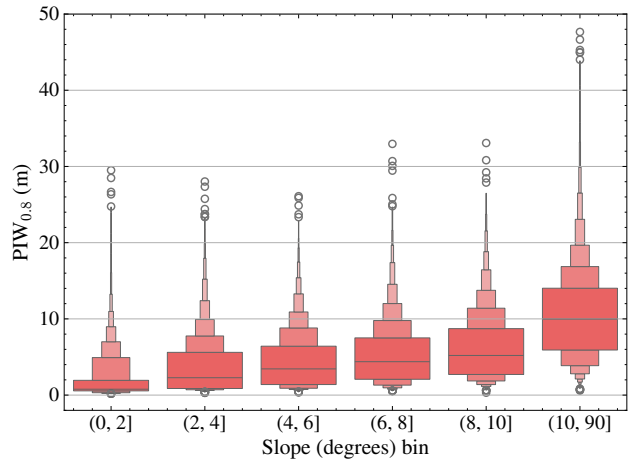


Figure 8: Boxplot of Prediction Interval Width at level $\alpha = 0.8$ in dependence of average slope within a 140 m surrounding.

tion offset at a forest border has a higher impact on the model than within a forest, where the spatial autocorrelation of tree heights is comparably high. Additionally, we observe an even higher uncertainty at forest borders, when the model is fine-tuned without the shift-resilient variant of the loss, see Appendix E for details.

4.2.7 Mountaineous Areas

As GEDI's footprint size is approximately 25m, measurements on steep slopes are overestimating true height, potentially having an estimated tree height for bare rocks. We therefore analyze the impact of slope on uncertainty by using the TandemX Edited Digital Elevation Model (EDEM) (González et al., 2020) to calculate the average slope in degrees in a 140 m surrounding, roughly corresponding to a 3×3 kernel for the EDEM. Figure 8 shows a boxplot of PIW in dependence on 6 different slope bins. There is a trend that with increasing slope, our model predicts an increase in PIW, as expected from higher label noise for steeper slopes for GEDI measurements (Fu et al., 2025). Note that all points with a PIW above 35 m are in areas with a slope of at least 10°. These findings indicate the need of effectively dealing with GEDI measurements in steep areas to reduce noise in the labels.

4.2.8 Qualitative Analysis

We qualitatively investigate uncertainty quantification based on one exemplary location, see Figure 9. An immediate observation is that the predictions of Lang et al.'s model are smoother than the predictions of our model. Except for this, the presented PIWs differ a lot. Overall, our model is more confident in its predictions compared to Lang et al.'s model. While the latter is less confident in areas with higher predictions, our model also shows confidence in some forest areas, see also Figure 1 again. The highest uncertainty for our model arises in the highest areas of the forest, and at forest borders. Therefore, the shape of the forest can also be read off by the uncertainty map of our model. An increase in uncertainty at forest borders cannot be observed in the predictions from Lang et al.'s model. Furthermore, note that our model is quite confident on grassland, where Lang et al.'s model returns higher uncertainty. Note that the model from Lang et al. is highly certain around the urban areas in the middle and the bottom of the image.

5 Conclusion

In this study, we propose a lightweight modification on pretrained models to incorporate uncertainty estimation, with only fine-tuning the added uncertainty head of the model. For this, we suggest using quantile regression, as it is known to perform well on image-to-image tasks in biology, see Angelopoulos et al. (2022).

This approach is experimentally tested in the application of tree height estimation, where a pretrained model from Pauls et al. (2025) is used to add and fine-tune the uncertainty head. Uncertainty estimates are compared to an existing approach from Lang et al. (2023). The results show that the uncertainty head can yield well-calibrated uncertainty quantification estimates after fine-tuning. The predicted quantiles are overall well-calibrated, although there are differences across target bins, as quantiles tend to underestimate tall trees. A general observation is that the model

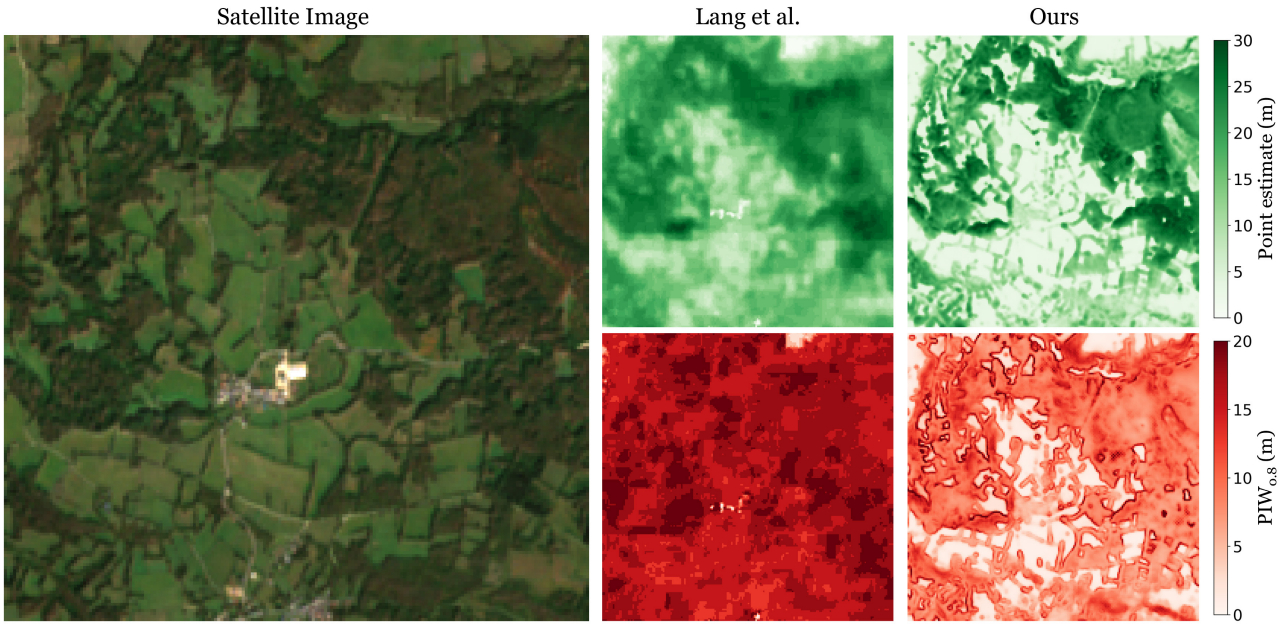


Figure 9: Sample satellite image with point estimates and 80% Prediction Interval Width of Lang et al.'s model and our model. Left: Satellite image (yearly median composite). Top Right: Point-estimates of canopy height. Bottom Right: 80% Prediction Inverval Width.

is less confident at forest borders and in steep terrain, hypothetically due to geolocation uncertainty in GEDI labels and label noise in mountainous areas. Another hypothesis is that wrong predictions with high confidence might be good indicators to detect faulty labels. A confirmation of this hypothesis and an eventual usage of detection and removal of faulty labels in training pipelines is a subject of future work.

All in all, the results show that well-calibrated uncertainty quantification is possible in the application area of tree height prediction based on satellite images. This is indispensable for non-linear downstream applications such as biomass and carbon storage estimation, as policy-making requires knowledge of worst-case scenarios and risk assessment.

Acknowledgements

This work was supported via the AI4Forest project, which is funded by the German Federal Ministry of Education and Research (BMBF; grant number 01IS23025A) and the French National Research Agency (ANR). We also acknowledge the computational resources provided by the PALMA II cluster at the University of Münster (subsidized by the DFG; INST 211/667-1).

References

Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., and Romano, Y. (2022). Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 717–730. PMLR.

Besic, N., Picard, N., Vega, C., Bontemps, J.-D., Hertzog, L., Renaud, J.-P., Fogel, F., Schwartz, M., Pellissier-Tanon, A., Destouet, G., et al. (2025). Remote-sensing-based forest canopy height mapping: some models are useful, but might they provide us with even more insights when combined? *Geoscientific Model Development*, 18(2):337–359.

Bian, C. and Xia, J. (2023). Uncertainty propagation in a global biogeochemical model driven by leaf area data. *Frontiers in Ecology and Evolution*, 11:1105832.

Candelas Bielza, J., Noordermeer, L., Næsset, E., Gobakken, T., Moan, M. Å., and Ørka, H. O. (2025). Assessing the effects of uncertainty in remote sensing-based predictions of tree species composition and site index on net present value calcula-

tions. *Forestry: An International Journal of Forest Research*, page cpaf057.

Dubayah, R., Armston, J., Healey, S. P., Bruening, J. M., Patterson, P. L., Kellner, J. R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z., Tang, H., Blair, J. B., Fatoyinbo, L., Goetz, S., Hancock, S., Hansen, M., Hofton, M., Hurtt, G., and Luthcke, S. (2022). GEDI launches a new era of biomass inference from space. *Environmental Research Letters*, 17(9):095001.

Fu, L., Shu, Q., Yang, Z., Xia, C., Zhang, X., Zhang, Y., Li, Z., and Li, S. (2025). Accuracy assessment of topography and forest canopy height in complex terrain conditions of southern china using icesat-2 and gedi data. *Frontiers in Plant Science*, 16:1547688.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(S1):1513–1589.

González, C., Bachmann, M., Bueso-Bello, J.-L., Rizzoli, P., and Zink, M. (2020). A fully automatic algorithm for editing the tandem-x global dem. *Remote Sensing*, 12(23).

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lang, N., Jetz, W., Schindler, K., and Wegner, J. D. (2023). A high-resolution canopy height model of the earth. *Nature Ecology and Evolution*, 7(11):1778–1789.

Lin, J., Gamarra, J. G., Drake, J. E., Cuchietti, A., and Yanai, R. D. (2023). Scaling up uncertainties in allometric models: how to see the forest, not the trees. *Forest ecology and management*, 537:120943.

Liu, S., Brandt, M., Nord-Larsen, T., Chave, J., Reiner, F., Lang, N., Tong, X., Ciais, P., Igel, C., Pascual, A., et al. (2023). The overlooked contribution of trees outside forests to tree cover and woody biomass across europe. *Science Advances*, 9(37):eadh4097.

Neumann, T. A., Martino, A. J., Markus, T., Bae, S., Bock, M. R., Brenner, A. C., Brunt, K. M., Cavanaugh, J., Fernandes, S. T., Hancock, D. W., et al. (2019). The ice, cloud, and land elevation satellite–2 mission: A global geolocated photon product derived from the advanced topographic laser altimeter system. *Remote sensing of environment*, 233:111325.

Nix, D. and Weigend, A. (1994). Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, pages 55–60 vol.1. IEEE.

Pan, Y., Birdsey, R. A., Phillips, O. L., Houghton, R. A., Fang, J., Kauppi, P. E., Keith, H., Kurz, W. A., Ito, A., Lewis, S. L., Nabuurs, G.-J., Shvidenko, A., Hashimoto, S., Lerink, B., Schepaschenko, D., Castanho, A., and Murdiyarso, D. (2024). The enduring world forest carbon sink. *Nature*, 631(8021):563–569.

Pauls, J., Zimmer, M., Kelly, U. M., Schwartz, M., Saatchi, S., CIAIS, P., Pokutta, S., Brandt, M., and Gieseke, F. (2024). Estimating canopy height at scale. In *Forty-first International Conference on Machine Learning*.

Pauls, J., Zimmer, M., Turan, B., Saatchi, S., CIAIS, P., Pokutta, S., and Gieseke, F. (2025). Capturing temporal dynamics in large-scale canopy tree height estimation. In *Forty-second International Conference on Machine Learning*.

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C. E., Armston, J., Dubayah, R., Blair, J. B., and Hofton, M. (2021). Mapping Global Forest Canopy Height through Integration of GEDI and Landsat Data. *Remote Sensing of Environment*, 253:112165.

Rolf, E., Gordon, L., Tambe, M., and Davies, A. (2024). Contrasting local and global modeling with machine learning and satellite data: A case study estimating tree canopy height in african savannas. *arXiv preprint arXiv:2411.14354*.

Schwartz, M., Ciais, P., De Truchis, A., Chave, J., Ottlé, C., Vega, C., Wigneron, J.-P., Nicolas, M., Jouaber, S., Liu, S., Brandt, M., and Fayad, I. (2023). 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.

Schwartz, M., Ciais, P., Ottlé, C., De Truchis, A., Vega, C., Fayad, I., Brandt, M., Fensholt, R., Baghdadi, N., Morneau, F., et al. (2022). High-resolution canopy height map in the landes forest (france) based on gedi, sentinel-1, and sentinel-2 data with a deep learning approach. *arXiv preprint arXiv:2212.10265*.

Smith, L. A. and Stern, N. (2011). Uncertainty in science and its role in climate policy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1956):4818–4841.

Steinwart, I. and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17.

Su, Y., Schwartz, M., Fayad, I., García, M., Zavala, M. A., Tijerín-Triviño, J., Astigarraga, J., Cruz-Alonso, V., Liu, S., Zhang, X., et al. (2025). Canopy height and biomass distribution across the forests of iberian peninsula. *Scientific Data*, 12(1):678.

Tolan, J., Yang, H.-I., Nosarzewski, B., Couairon, G., Vo, H. V., Brandt, J., Spore, J., Majumdar, S., Haziza, D., Vamaraju, J., Moutakanni, T., Bojanowski, P., Johns, T., White, B., Tiecke, T., and Couprie, C. (2024). 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.

Yambayamba, A. M., Handavu, F., Kapinga, K., and Jucker, T. (2025). Tree height uncertainty biases aboveground biomass estimation more than wood density in miombo woodlands. *EGUsphere*, 2025:1–37.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   Not Applicable
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   Not Applicable
   (b) Complete proofs of all theoretical results.
   Not Applicable
   (c) Clear explanations of any assumptions.
   Not Applicable

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   No, training dataset is too big to share (about 50 TB). All code to deduce metrics/figures from result dataframes are included in Supplemental Material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   No, cluster type will be included upon acceptance, as it also has to be included in acknowledgement.

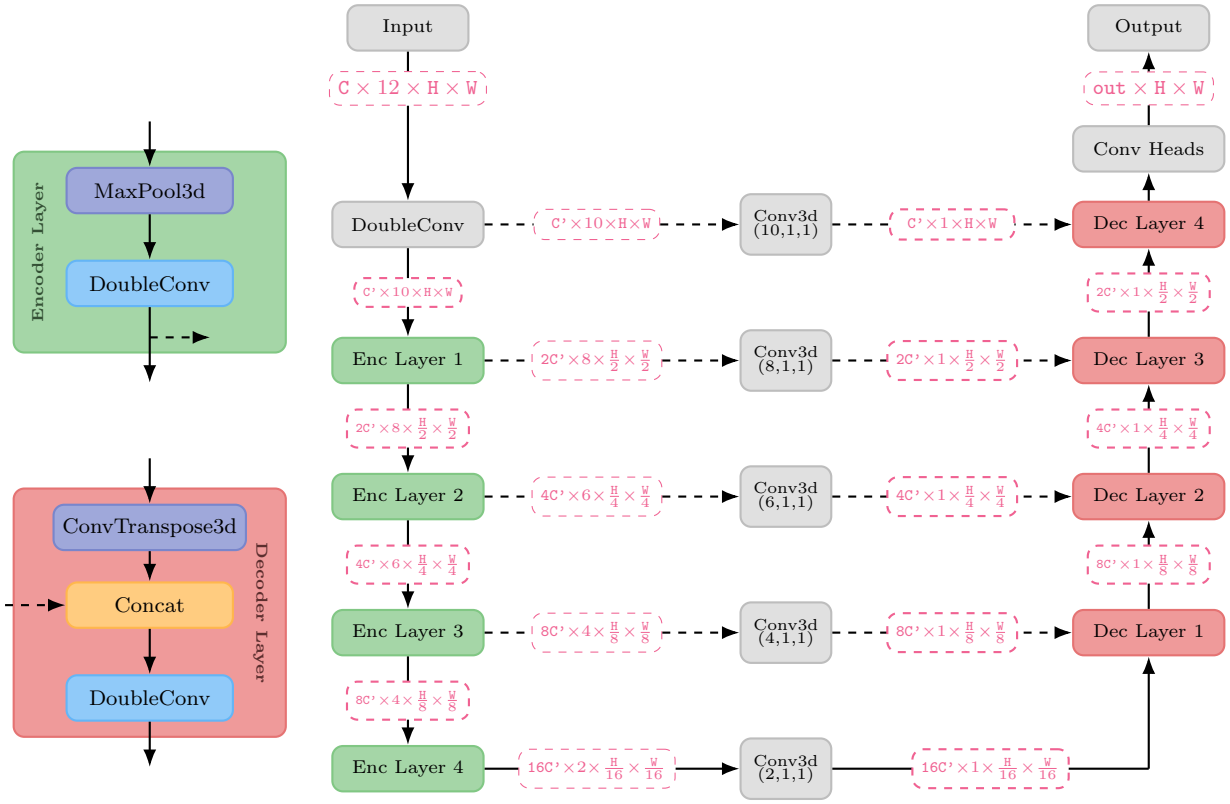
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   Yes
   (b) The license information of the assets, if applicable.
   Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable.
   Not Applicable
   (d) Information about consent from data providers/curators.
   Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
   Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.
   Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
   Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
   Not Applicable

Canopy Tree Height Estimation Using Quantile Regression: Modeling and Evaluating Uncertainty in Remote Sensing Supplementary Materials

A Model Architecture

The used architecture is depicted in Figure 10. Please note that the encoder and decoder originates in the 3d-Unet from Pauls et al. (2025), while one convolutional head is added to predict 10 quantiles. In particular, $C = 16$, $H = W = 256$, $C' = 64$ and out $= 11$ in our model.



(a) Layer details (b) 3D UNet architecture with Temporal Aggregation

Figure 10: Architecture of the 3D UNet. (a) Each Encoder Layer applies MaxPool3d for 2× spatial downsampling followed by a DoubleConv block (two Conv3d + GroupNorm + ReLU); the initial DoubleConv omits the MaxPool. Each Decoder Layer applies ConvTranspose3d for 2× spatial upsampling, concatenates skip features (dashed), then applies a DoubleConv block. (b) Overall architecture. Encoder features pass through Temporal Aggregation Conv3d layers with temporal-only kernels (T,1,1) that collapse the time dimension to 1 before entering the decoder as skip connections. The solid path at the bottom shows the bottleneck flow. Tensor shapes are shown in magenta as $C \times T \times H \times W$.

B Ablation study with Gaussian Regression and Log-Gaussian Regression

As an ablation study, we compare our quantile regression approach against two alternative uncertainty estimation methods: Gaussian Regression and Log-Gaussian Regression. In Gaussian Regression, the model predicts, for each pixel, a mean μ and a log-variance $\log \sigma^2$, and is trained by minimizing the Gaussian negative log-likelihood:

$$\mathcal{L}(\mu, \sigma^2, y) = \frac{1}{2} \log \sigma^2 + \frac{(y-\mu)^2}{2\sigma^2},$$

following the same formulation as Lang et al. (2023). Compared to our architecture, this requires only one additional output channel (for $\log \sigma^2$) instead of ten (for the quantile predictions), resulting in comparable computational costs. A limitation of Gaussian Regression is that the symmetric Gaussian distribution assigns non-zero probability to negative tree heights. To address this, we additionally consider Log-Gaussian Regression, which replaces the Gaussian likelihood with a log-normal likelihood. In this formulation, the model predicts the parameters of a Gaussian distribution in log-space, ensuring that the implied distribution over tree heights has support strictly on the positive real line. The network architecture remains unchanged from Gaussian Regression. All models are trained with and without the Shift-Resilient Loss (see Section 3.3).

Table 3 and Table 4 report the same point estimation and uncertainty metrics as in the main paper for all models tested.

Table 3: Results on point-estimators, measured using Mean Squared Error (MSE, m^2), Mean Absolute Error (MAE, m), R^2 and Empirical Coverage (EC).

Model	MSE ↓	MAE ↓	R2 ↑	$EC_{0.5}$
Lang et. al.	38.32	4.25	0.55	0.47
Pauls et. al.	20.64	1.90	0.76	0.58
Gaussian w/o Shift	22.11	2.17	0.74	0.58
Gaussian w/ Shift	22.54	2.17	0.74	0.58
Log Gaussian w/o Shift	21.86	2.05	0.74	0.54
Log Gaussian w/ Shift	21.86	2.04	0.74	0.54
Ours wo/ Shift	20.34	1.88	0.76	0.50
Ours	20.81	1.90	0.76	0.50

Table 4: Comparison of Mean Prediction Interval Width (MPIW) and Prediction Interval Coverage Probability (PICP) for multiple uncertainty levels α.

Metric	MPIW					PICP				
α	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
Lang et. al.	5.54	6.87	8.39	10.22	12.76	0.37	0.45	0.52	0.58	0.64
Gaussian w/o Shift	3.93	4.91	6.04	7.47	9.59	0.70	0.78	0.85	0.90	0.94
Gaussian w/ Shift	3.79	4.72	5.82	7.19	9.23	0.66	0.75	0.82	0.88	0.93
Log Gaussian w/o Shift	3.67	4.62	5.76	7.24	9.61	0.63	0.74	0.82	0.89	0.94
Log Gaussian w/ Shift	3.35	4.22	5.25	6.59	8.73	0.58	0.69	0.79	0.86	0.92
Ours w/o Shift	2.69	3.34	4.15	5.25	7.10	0.49	0.59	0.69	0.79	0.90
Ours	2.43	3.01	3.72	4.68	6.34	0.47	0.56	0.66	0.76	0.87

C Computational Overhead

In this section we provide the results of a comparison between our proposed method and the one, which we build upon (Pauls et al., 2025). It can be seen that the additional uncertainty head adds less than 10% of extra time and less than 13% additional VRAM usage during inference.

Table 5: VRAM usage and compute time during inference and training for a single input, FLOPs and number of parameters from our proposed method and Pauls et al. (2025), both with either frozen backbone and trainable head (Ours and Pauls et al.$_{head_only}$) or fully trainable (Ours$_{full}$ and Pauls et al.$_{full}$).

Model	VRAM		Time (ms)		FLOPS	Params
	Inference	Training	Inference	Training	(G)	(M)
Ours	1668MiB	1672MiB	2.79 ± 0.72	3.82 ± 0.18	430.531	85.0659
Ours$_{full}$	1668MiB	3810MiB	2.79 ± 0.72	7.54 ± 0.25	430.531	85.0659
Pauls et al.$_{head_only}$	1476MiB	1478MiB	2.55 ± 0.97	3.00 ± 0.17	425.623	85.0653
Pauls et al.$_{full}$	1476MiB	3900MiB	2.55 ± 0.97	6.77 ± 0.26	425.623	85.0653

D Empirical Coverage across Prediction Bins

The empirical coverage of quantile predictions across predictions bins is depicted in Figure 11.

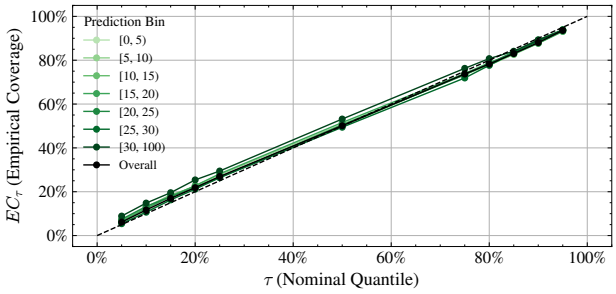


Figure 11: Empirical Coverage vs Nominal Quantile for multiple prediction bins for our model. Across all prediction bins, the conditional quantiles are well-calibrated.

E Visualizations for our Model without Shifted Loss

Figure 12 shows the EC for different quantiles on the left and split up by target bins on the right side, Figure 13 shows the boxenplots for forest borders and different slope values and Figure 14 shows a visual comparison between training our model with and without shifted loss, showing that the model without shift loss is much less confident at forest borders.

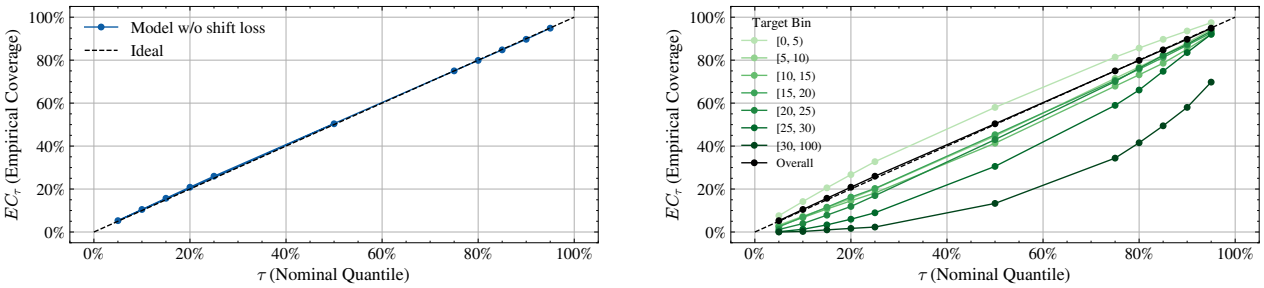


Figure 12: Empirical Coverage vs Nominal Quantile for our model trained without shift loss.

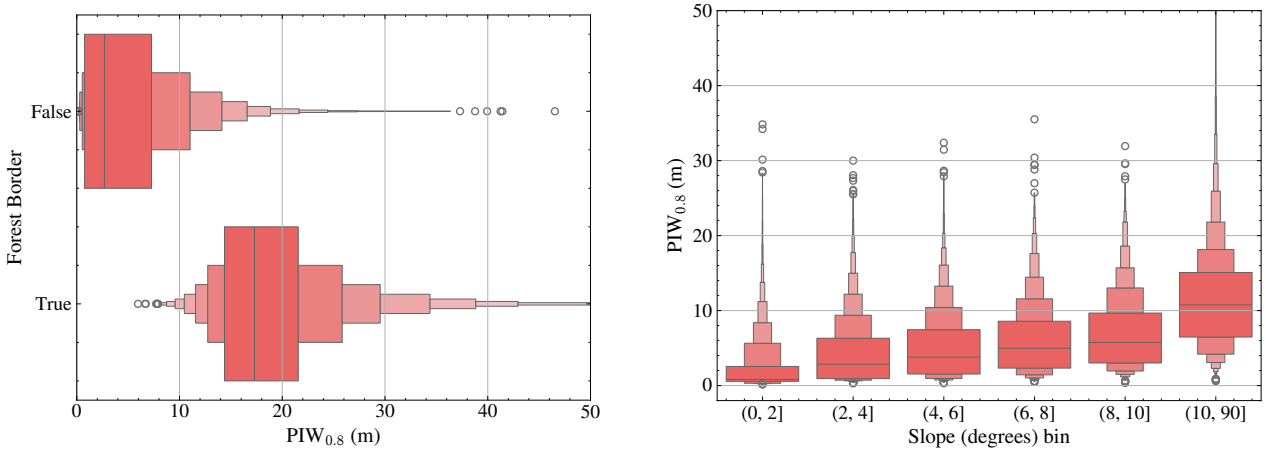


Figure 13: Boxplot of Prediction Interval Width at level $\alpha = 0.8$ for model trained without shift loss in dependence of either Left: being at forest border or Right: average slope within a 140 m square. Note that the model without shift loss is more uncertain at forest borders compared to the model with shift loss (compare Figures 7 and 8).

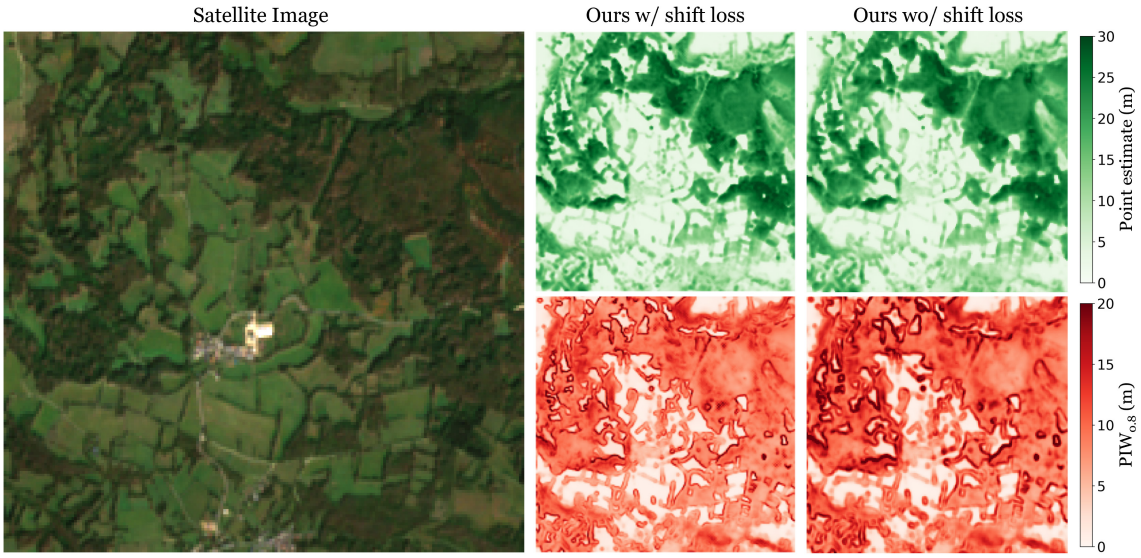


Figure 14: Sample satellite image with point estimates and 80% Prediction Interval Width of our model, trained with or without shift loss. Left: Satellite image (yearly median composite). Top Right: Point-estimates of canopy height. Bottom Right: 80% Prediction Inverval Width. It is clearly visible that the shift loss reduces uncertainty at forest borders.