ISPRS Journal of Photogrammetry and Remote Sensing 195 (2023) 269–286

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Country-wide retrieval of forest structure from optical and SAR satellite imagery with deep ensembles

Alexander Becker [a,*], Stefania Russo [a], Stefano Puliti [b], Nico Lang [a], Konrad Schindler [a], Jan Dirk Wegner [a,c]

[a] EcoVision Lab, Photogrammetry and Remote Sensing, ETH Zurich, Switzerland
[b] National Forest Inventory Department, Norwegian Institute of Bioeconomy Research (NIBIO), Norway
[c] Institute for Computational Science, University of Zurich, Switzerland

ARTICLE INFO

Keywords:
Bayesian deep learning
Forest structure
Multispectral
SAR
Country-scale
Sentinel

ABSTRACT

Monitoring and managing Earth's forests in an informed manner is an important requirement for addressing challenges like biodiversity loss and climate change. While traditional in situ or aerial campaigns for forest assessments provide accurate data for analysis at regional level, scaling them to entire countries and beyond with high temporal resolution is hardly possible. In this work, we propose a method based on deep ensembles that densely estimates forest structure variables at country-scale with 10-m resolution, using freely available satellite imagery as input. Our method jointly transforms Sentinel-2 optical images and Sentinel-1 synthetic-aperture radar images into maps of five different forest structure variables: 95th height percentile, mean height, density, Gini coefficient, and fractional cover. We train and test our model on reference data from 41 airborne laser scanning missions across Norway and demonstrate that it is able to generalize to unseen test regions, achieving normalized mean absolute errors between 11% and 15%, depending on the variable. Our work is also the first to propose a variant of so-called Bayesian deep learning to densely predict multiple forest structure variables with well-calibrated uncertainty estimates from satellite imagery. The uncertainty information increases the trustworthiness of the model and its suitability for downstream tasks that require reliable confidence estimates as a basis for decision making. We present an extensive set of experiments to validate the accuracy of the predicted maps as well as the quality of the predicted uncertainties. To demonstrate scalability, we provide Norway-wide maps for the five forest structure variables.

1. Introduction

Forest structure relates to the three-dimensional (3D) spatial arrangement of the plant community within a forest and is both a result and a driver of ecosystem processes and biological diversity (Spies, 1998). Forest structure is strongly correlated with a forest's ability to store carbon and to provide habitat for a variety of species (Turner et al., 2003; Bergen et al., 2009; Dubayah et al., 2020). Given the large extent and dynamic nature of forest ecosystems, it is important to develop ways of mapping and monitoring their structure consistently through space and time (Valbuena et al., 2020). Such maps can support a more informed and dynamic management of forest resources, required to tackle important global challenges such as the loss of biodiversity; and the mitigation of, as well as adaptation to climate change.

Measuring forest structure has long been an expensive and time-consuming task. Traditional in-situ methods are limited to the recording of a few tree characteristics measurable from the ground (e.g., diameter at breast height, tree species, to some degree tree height), and restricted to small sample plots sparsely scattered across the landscape. Consequently, in-situ measurements are not sufficient for a spatially explicit understanding of forest structures. To collect additional, explicit forest structure measurements, field direct observations can be complemented with terrestrial laser scanning (TLS). Still, TLS in practice only yields local samples, as it has a limited range and is affected by occlusions (Calders et al., 2020). In contrast, airborne laser scanning (ALS) allows one to densely observe forest structure at regional scales. Forest structural variables derived from ALS, such as canopy height, cover, or density, have long been used for the wall-to-wall characterization of single-tree (Hyyppä et al., 2001) and forest biophysical properties (Nilsson, 1996; Næsset, 2002). But the high operational cost, which scales more or less linearly with flight time, limits the coverage and

* Corresponding author.
E-mail address: alexander.becker@geod.baug.ethz.ch (A. Becker).

https://doi.org/10.1016/j.isprsjprs.2022.11.011
Received 19 January 2022; Received in revised form 2 June 2022; Accepted 18 November 2022
Available online 7 December 2022
0924-2716/© 2022 The Author(s). Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

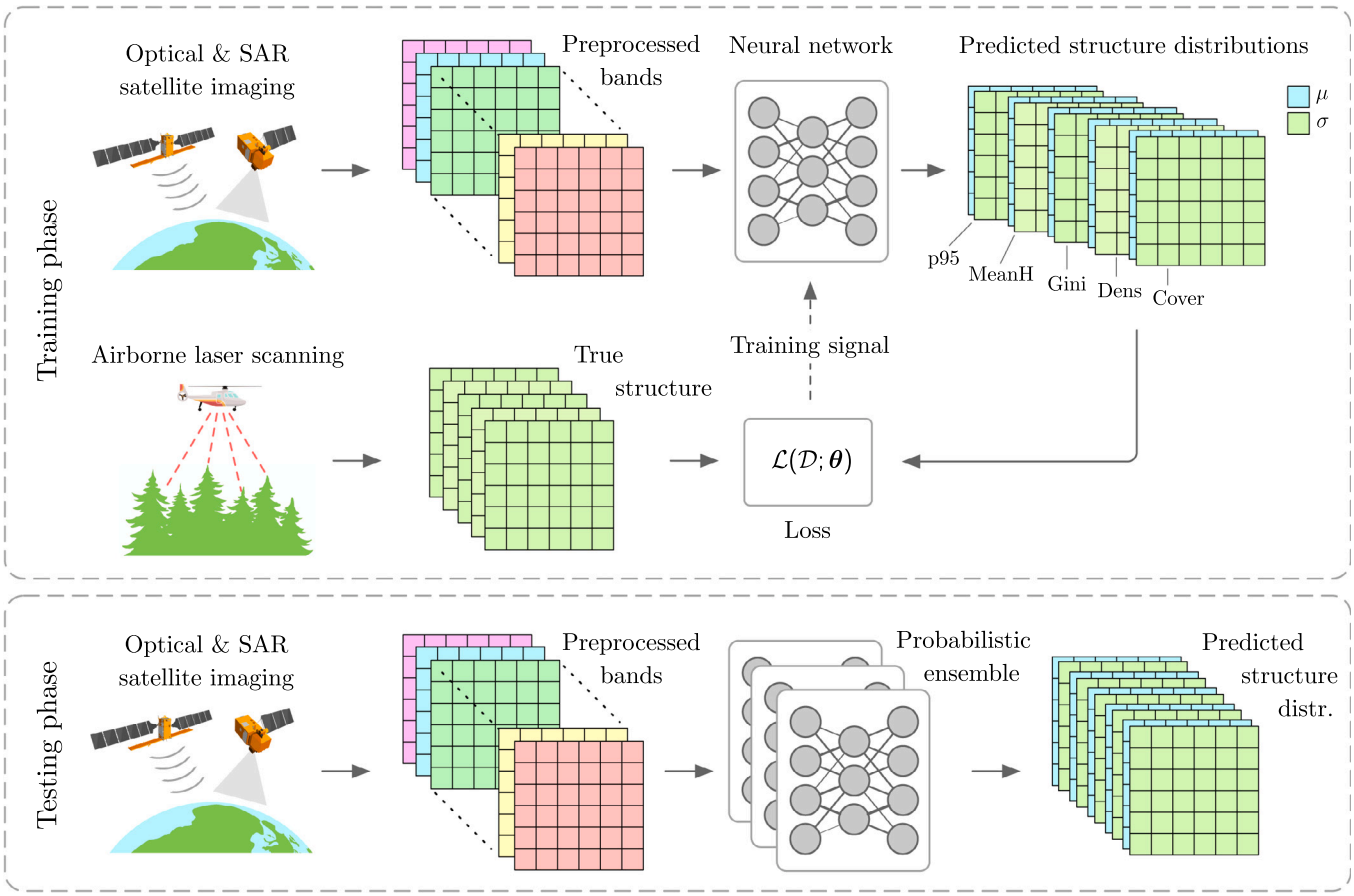

Fig. 1. High-level overview of our method. *Top panel*: During training, the parameters of a neural network model are optimized to reduce the deviation between predicted and ALS-derived forest structure variables. *Bottom panel*: During inference, an *ensemble* of multiple neural networks predicts a distribution over forest structure variables, given only optical and SAR satellite images with 10 m GSD as input. Satellite icons from FU Berlin (2019).

revisit time of ALS campaigns and is a bottleneck for country-scale applications.

The natural next step to scale beyond local and single-date ALS campaigns is to take advantage of the wealth of open data from Earth observation satellites. A prominent example is the European Union's Copernicus program, with multiple satellite missions whose data products are useful to predict forest structure dynamics (Puliti et al., 2021). In particular, the Sentinel-1 (synthetic-aperture radar, SAR) and Sentinel-2 (multispectral optical) missions provide near-global coverage, with both high spatial resolution (~10 m ground sampling distance, GSD) and frequent revisit times (≤5 days; the effective temporal resolution of the optical sensor is reduced by the cloud coverage). These missions deliver observations over a wide spectral range at an unprecedented rate, but can they be converted to forest structure products? A new challenge arises to retrieve the relevant information from multispectral image data at ~10 m GSD, rather than LiDAR waveforms with <1 m^2 footprints.

In the absence of a tractable radiative transfer model, supervised machine learning offers a statistical approach to retrieve environmental parameters from image data. In recent years, deep neural networks (DNNs) have emerged as the mainstream tool for image analysis, including their applications in remote sensing and, in particular, mapping of forest structure (Lang et al., 2019). DNNs have brought substantial performance gains through their ability to learn the complete functional mapping from raw image values to the desired output variables. While DNNs often achieve high precision, their outputs are typically limited to point estimates and/or are overly confident (Guo et al., 2017; Gast and Roth, 2018). As well-calibrated estimates of uncertainty are important whenever the model outputs feed into critical decisions or probabilistic models, the Bayesian interpretation of model outputs is an active line of research (Blundell et al., 2015; Gal and Ghahramani, 2016; Guo et al., 2017; Kendall and Gal, 2017; Lakshminarayanan et al., 2017; Ovadia et al., 2019; Gustafsson et al., 2020; Wilson and Izmailov, 2020).

In this paper, we go a step further and exploit recent advances from the field of *Bayesian deep learning* (BDL) to densely map forest structure variables *and* associated uncertainties from optical and SAR satellite images. BDL is understood as the deep learning counterpart to traditional Bayesian inference techniques, for instance those for linear regression (Rasmussen, 2004, Ch. 2 to 2.1.1) or Gaussian process regression (Rasmussen, 2004, Ch. 2.1.2 to 2.3). Instead of committing to a single solution of model parameters, BDL is characterized by (approximate) marginalization over a posterior distribution of all possible models, given some assumptions about the prior and the data likelihood. Such a principled approach is particularly attractive for data-driven models, where the model uncertainty is not a consequence of explicit modeling decisions, but instead is caused by the limited training data. Bayesian marginalization may improve the predictive accuracy, but more importantly it can give the user access to well-calibrated estimates of the predictive uncertainty, which take into consideration how well a given model prediction is supported by training data. This ability to self-diagnose the reliability of each individual prediction is indispensable for many downstream tasks that ingest the model output as their input. As an example, when making informed decisions, it is crucial to know how reliable the data are that the decisions are based on. Uncertainty estimates have been successfully used for decision making in many fields of application (Council, 2006; Soroudi and Amraee, 2013; Martin and Johnson, 2019; Sniazhko, 2019;

Alzate-Mejía et al., 2021), and we argue that they are equally relevant in forest management in order to optimize decisions for cost, carbon stock, biodiversity etc.

Our method (see overview in Fig. 1) is fully supervised and consists of a training and a testing phase. During the training phase, a deep convolutional neural network is optimized on training data, i.e., optical and SAR images with associated, pixel-wise forest structure reference values computed from ALS data. The fitting employs a loss function that does not just penalize deviations between model outputs and ALS reference data. Instead, the loss is proportional to the negative log posterior probability over the model parameters given the data, assuming a zero-mean normally distributed prior (acting as a regularizer) and a Gaussian data likelihood. Thus, for each pixel and variable, the model outputs the parameters of a Gaussian distribution, thereby capturing uncertainty inherent in the input data (aleatoric uncertainty). During the testing phase, multiple trained networks are combined into an ensemble model, exploiting the stochastic nature of neural network initialization and training. The predicted distributions of all networks are aggregated into an ensemble estimate of the distribution over structure variables. The aggregation of predictions from independently trained models corresponds to a sample-based approximation of Bayesian marginalization (Gustafsson et al., 2020; Wilson and Izmailov, 2020). It captures the uncertainty in the model parameters (epistemic uncertainty), which is combined with the aleatoric uncertainty to obtain the overall uncertainty of the predicted forest structure variables.

In this work, we propose a deep ensembles approach to predict structural forest variables often used in modeling forest biophysical properties (95th height percentile, mean height, density, Gini coefficient and fractional cover), using optical and SAR satellite images as input. Our approach can be seen as a scalable complement to regional ALS, enabling forest structure mapping with 10-m ground sampling distance (GSD) at country-scale and with a high update frequency (in the extreme case down to five days), at a low cost. Moreover, the estimated forest structure maps come with an individual, calibrated uncertainty estimate for every structure variable at every single pixel. We conduct extensive experiments on a test region in Norway, for which reference data from a full-waveform ALS campaign is available. We then apply our method to compute a country-wide forest structure map for all of Norway, that we make available online. Source code for training and testing the model is provided, too.

2. Related work

2.1. Remote sensing of structural forest variables

Since its early days, ALS research has seen a tremendous growth, and forest structural variables derived from ALS are broadly used as a source of auxiliary information for the large scale characterization of forest ecosystems. Many different types of modeling techniques have been tested over the years, most of which rely on linking field measured forest biophysical properties (e.g. biomass) with forest structural predictor variables derived from ALS, like height percentiles, vegetation density, forest cover, or foliage height diversity (Næsset et al., 2004; Næsset, 2007). More recently, there has been a trend to directly use ALS-derived structural variables to map forest structural (Coops et al., 2016; Valbuena et al., 2017; Adnan et al., 2021) and functional diversity (Schneider et al., 2017; Zheng et al., 2021). While ALS data are the most detailed source of information to characterize forest structure, their geographic and temporal availability remains limited.

If the goal is frequent, or even continuous, monitoring of large forest areas, a more promising data source are freely available satellite images, such as those provided by the Landsat and Copernicus programs. Landsat-based approaches have mostly been relying on time series features to map forest structure (Tyukavina et al., 2015; Hansen et al., 2016; Potapov et al., 2019, 2021). Also Copernicus' Sentinel-1 and Sentinel-2 missions, with 10 m GSD and less than five days revisit time, offer dense time series of observations that are suitable for forest monitoring. Previous studies have demonstrated the usefulness of Sentinel data to map and estimate key forest biophysical variables (e.g., above ground biomass) (Laurin et al., 2018; Puliti et al., 2020; Breidenbach et al., 2021) and their dynamics trough time (Puliti et al., 2021). In the past two years, a growing body of literature has shown the possibility to map forest canopy height with Sentinel-2 (Lang et al., 2019; Shimizu et al., 2020; Astola et al., 2021; Lang et al., 2021). Since canopy height is only one of many ecosystem characteristics that can be obtained from ALS, our work aims to understand whether Sentinel data can provide maps for a more comprehensive spectrum of structural variables, like vegetation density, cover, and complexity. Like (Lang et al., 2019) our model learns to extract predictive texture features from single input images.

2.2. Deep learning in remote sensing

In the last decade, deep learning has revolutionized the way information is extracted from images. In particular, convolutional neural networks (CNNs) have achieved unprecedented results in areas like image classification (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; He et al., 2016), semantic segmentation (Long et al., 2015; Chen et al., 2016), object detection (Szegedy et al., 2013; Girshick et al., 2014; Redmon et al., 2016) and further perception tasks.

In addition, deep learning is increasingly being applied to vision tasks in remote sensing, such as super resolution (Lanaras et al., 2018; de Lutio et al., 2019) or change detection (Caye Daudt et al., 2018). Traditional applications include land cover classification from aerial (Kaiser et al., 2017; Marmanis et al., 2018; Zhang et al., 2019) or satellite images (Helber et al., 2019). Agricultural crop type classification, a specific type of land cover, is a well-studied task, where most authors exploit features from time-series data (Rußwurm and Körner, 2018; Rustowicz et al., 2019; Garnot et al., 2019; Rußwurm and Körner, 2020; Turkoglu et al., 2021a,b). More related to our work are methods that seek to predict biophysical indicators such as crop yield from climate data and Enhanced Vegetation Index maps (Kuwata and Shibasaki, 2015), the prediction of sea ice concentration from RADARSAT SAR satellite images (Wang et al., 2016) or tree density estimation from Sentinel-2 optical images (Rodríguez and Wegner, 2018; Rodríguez et al., 2021). Most works so far do not accompany their map products with calibrated uncertainty estimates. Exceptions include (Rodríguez et al., 2021), where uncertainties are used to guide active learning; and Lang et al. (2022), where BDL is integrated into a 1-dimensional CNN for spaceborne LiDAR analysis. The prior work most relevant for our paper is Lang et al. (2019), where a CNN was developed to predict country-wide vegetation height maps from Sentinel-2 optical images. Their model was trained and evaluated for Gabon and Switzerland, where training data were derived from LiDAR measurements and photogrammetric surface reconstruction, respectively. Here, we predict five forest structural variables jointly and estimate their predictive uncertainties using an ensemble of probabilistic neural networks. Moreover, we explore the potential of data fusion using Sentinel-1 SAR images as an additional input signal on top of the Sentinel-2 optical bands.

2.3. Uncertainty in deep learning

Although deep learning has become the most widely used method in computer vision, model outputs are often trusted "blindly". In regression problems, outputs are usually point estimates with no attached notion of uncertainty, while classification scores have been shown to be overconfident (Guo et al., 2017; Lakshminarayanan et al., 2017). To mitigate this effect and to develop more trustworthy models, reliably quantifying the predictive uncertainty is important -and an open research question (Gustafsson et al., 2020).

Usually, uncertainty in machine learning is decomposed into *aleatoric* and *epistemic* uncertainty (Gal and Ghahramani, 2016). The former is assumed to be inherent in the observations – e.g. resulting from sensor noise or lack of signal – and can thus not be "explained away" with more training data. On the contrary, epistemic uncertainty captures uncertainty in the model itself, i.e., the lack of knowledge due to not having seen enough training data for the respective region of the input space. Unlike aleatoric uncertainty, it cannot simply be learned from data — instead, it generally requires marginalization over an (approximate) posterior distribution over model parameters. These methods are colloquially referred to as *Bayesian deep learning* (Kendall and Gal, 2017); common approaches include variational Bayesian inference (Ranganath et al., 2014; Blundell et al., 2015) and methods based on Markov chain Monte Carlo (MCMC) sampling (Neal, 1996; Welling and Teh, 2011; Chen et al., 2014). In this work, we use a deep ensemble (Lakshminarayanan et al., 2017), a method specifically developed for deep neural networks that can be understood to perform approximate Bayesian inference (Gustafsson et al., 2020). The general idea is to train an ensemble of M independent models on the same data, each initialized with a different set of random weights. The randomness inherent in the weight initialization, as well as random sampling of training batches, causes each model to converge to a different local minimum in the solution space, and the resulting weights can be interpreted as samples from an approximate posterior distribution (Gustafsson et al., 2020; Wilson and Izmailov, 2020). In practice, among all methods performing approximate Bayesian inference in deep learning, ensembles are generally reported to achieve the best results in terms of predictive performance and reliability of the produced uncertainty estimates (Ovadia et al., 2019; Gustafsson et al., 2020; Ashukha et al., 2020).

3. Data

3.1. ALS structural forest variables

In this study, the target variables are forest structural variables normally extracted from ALS data. In Norway, the corresponding ALS data are available as part of a national program aimed at the production of a high resolution digital terrain model (DTM) of the entire country. Amongst the many available ALS projects, we selected a sample of 41 projects covering a range of latitude (58°N – 69°N), longitude (5°E-18°E), and the corresponding diversity in landscapes and forest types in Norway. Our selected areas are shown in Fig. 2 and are covered by the following main forest types:

- East: productive boreal forest in the south eastern part, characterized by mild slopes and continental climate.
- North: predominantly deciduous areas with a large portion of low-productivity mountain birch (*Betula nana*) in the north. These forests represent the transition from subpolar oceanic vegetation towards the tundra and sub-arctic climates.
- West: deciduous forest alternating with patches of coniferous forests and un-productive forests in the coastal areas. These areas are characterized by a milder oceanic climate that, due to the steep slopes (i.e., fjords), can transition to tundra and alpine vegetation within short distances.

The selected samples of ALS data were collected between 2015 and 2018, with ~80% being from the 2016–2017 period. For ~35% of the area, the ALS data were collected under leaf-off conditions (Oct.-Dec.). The point density in the selected ALS projects was 2 pts./m^2 for 51% of the area, 5 pts./m^2 for a further 48%, and 10 pts./m^2 for the remaining 1%. Overall, the sample covers a broad range of climatic, vegetation, terrain, and ALS data characteristics, thus providing a suitable training set for deep learning models that shall generalize across a range of conditions.

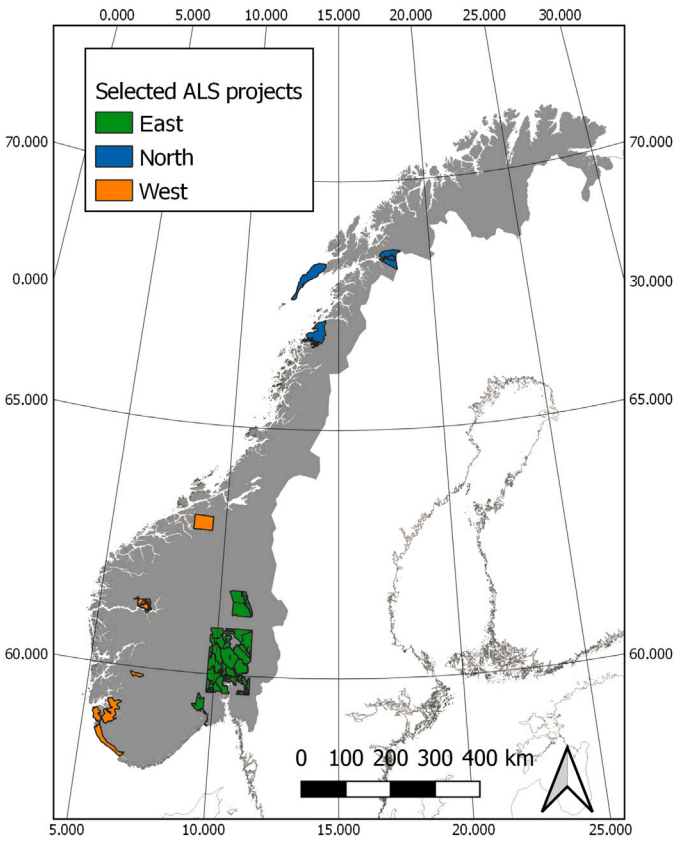

Fig. 2. Overview of the selected ALS project used in this study for training, validation, and testing of the developed methods. The ALS projects are colored based on their geographical grouping. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The height values z (in meters above sea level) from the raw ALS point clouds were normalized to Dz (in meters above ground) by subtracting the ground elevation from each point's z value. Furthermore, all points with $Dz > 1.3$ m were classified as vegetation points. We extracted a selection of commonly used ALS variables to describe forest structure and ultimately ecosystem characteristics (Valbuena et al., 2020): height, density/cover, complexity, and habitat area. These variables were selected with the aim to minimize the number of variables while maintaining complementary information on the vertical and horizontal distribution of the forest canopy. While these variables do not provide direct quantitative biophysical properties used in traditional forest management inventories (e.g. above ground biomass), they remain useful proxies for characterizing for forest monitoring purposes (Senf and Seidl, 2022; Hansen et al., 2013). The variables were computed from the normalized point clouds (i.e., m above ground) on a 10 m raster, which was then bi-linearly resampled to match the Sentinel-2 pixels. The following variables were computed for the entire area where ALS was available. We show corresponding histograms in Fig. 3 and an illustration of the variables in Fig. 4.

- P95 (m above ground): the 95th percentile of the Dz values of all vegetation points. This variable is close to the maximum height but it removes potential noise from spurious high ALS returns (e.g., from birds or power lines). The use of the height percentiles dates back to the some of the first area-based forest inventories (Næsset, 2002). Among them, the 95th percentile is useful as a measure of canopy top height, which is known to be correlated with the developmental stage of the forest and hence its biomass stock. Furthermore, canopy height diversity has been shown to be useful to monitor forest disturbance regimes (Senf et al., 2020).

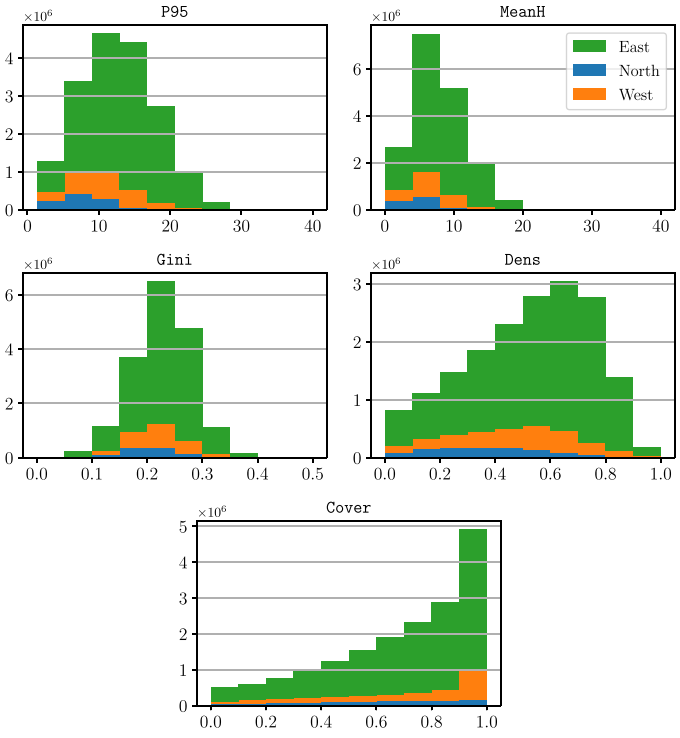

Fig. 3. Histograms of the five ALS-derived structural forest variables. The coloring indicates the different geographical areas, and bars are stacked such that the overall bar height denotes the overall number of data points (i.e., pixels) for a given bin. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

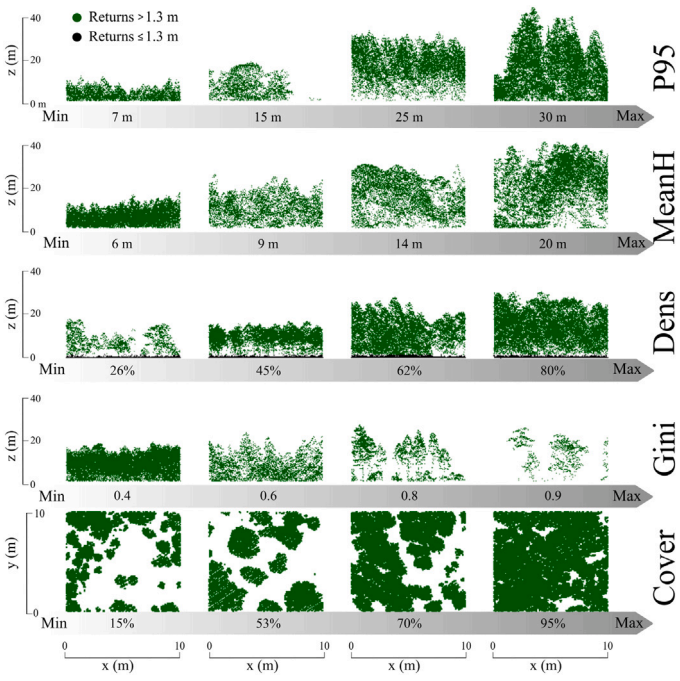

Fig. 4. Visual representation of the variability in forest structures captured in the range of the studied ALS forest structural variables, including the 95th percentile (P95; m), mean height (`MeanH`; m), vegetation density (`Dens`), Gini coefficient of the Dz values (`Gini`, %), and forest cover (`Cover`; %). For P95, `MeanH`, `Dens`, and `Gini` the scatter plots represent a side view of the normalized point clouds corresponding to Sentinel-2 pixels, while for the Cover the plots represent top-down view of plots representing the canopy a (in green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- `MeanH` (m above ground): mean of the Dz values of vegetation points. This variable is of interest because it not only captures the height of a forest but also the vertical distribution of the plant material within the canopy. We note that, thanks to its ability to encompass these two levels of information, `MeanH` is at present the only predictor variable used to produce country-wide maps of forest biomass in Norway (Astrup et al., 2019).
- `Dens` (%): proportion of vegetation points in the entire set of LiDAR returns. This variable describes the density of the forest canopy and, along with the height percentiles, has long been used in area-based ALS forest inventories. Its is complementary to the canopy height, as it provides information about the vertical distribution of the plant material through the canopy (Næsset, 2002).
- `Gini` (index): The Gini coefficient is equal to half of the relative mean difference in Dz values among all the vegetation returns, and was calculated using the function implemented in the *leafR* package (Alves de Almeida et al., 2020). The Gini coefficient is a measure of the inequality among the members of a data distribution, and it has been used as a proxy for tree size variation (Knox et al., 1989) and to map differences in forest structures and management regimes (Valbuena et al., 2017; Adnan et al., 2019). While typically the Gini coefficient has been calculated using single-tree data, a recent work by Adnan et al. (2021) demonstrated the usefulness of the Gini coefficient calculated from the ALS Dz values, showing that it could reliably describe the structural heterogeneity of the forest.
- `Cover` (%): Forest cover in terms of the proportion of projected canopy area relative to the entire area of a pixel (100 m^2). Cover was computed by projecting the vegetation points onto a (x, y) plane and converting them to a binary occupancy grid with 1 m resolution. The forest cover was then derived as the percentage of pixels occupied by forest. While somewhat correlated to `Dens`, cover remains a fully 2-dimensional variable, describing the horizontal vegetation cover, rather than the density of points in the vertical canopy profile. Time series of forest cover maps have been widely used in remote sensing as a measure of the canopy openness, and to assess land use changes with particular interest on forest cover losses (Hansen et al., 2013) and forest disturbance (Senf et al., 2018).

We divide each ALS project geographically into horizontal stripes of width 900 pixels (9.0 km). Within each stripe we assign the northernmost 5.4 km to the training set, the next 1.8 km to the validation set and the southernmost 1.8 km to the test set. In this way, the different regions and modalities in each ALS area are evenly distributed between the three sets. Overall, the data set consists of 105,022,419 pixels (10,502 km^2) of ALS reference data, divided into 64,487,551 training pixels (6449 km^2), 20,784,407 validation pixels (2078 km^2) and 19,750,461 test pixels (1975 km^2), following the "60-20-20" ratio often used in machine learning studies.

3.2. Sentinel satellite imagery

Sentinel-1 and Sentinel-2 are satellite missions that belong to the European Union's Copernicus Earth observation programme (European Space Agency, 2021a), providing high-resolution SAR and optical imagery of land and coastal water areas between 56° South and 84° North. Each mission consists of a ground segment and a constellation of two satellites in sun-synchronous low earth orbits, phased 180° from

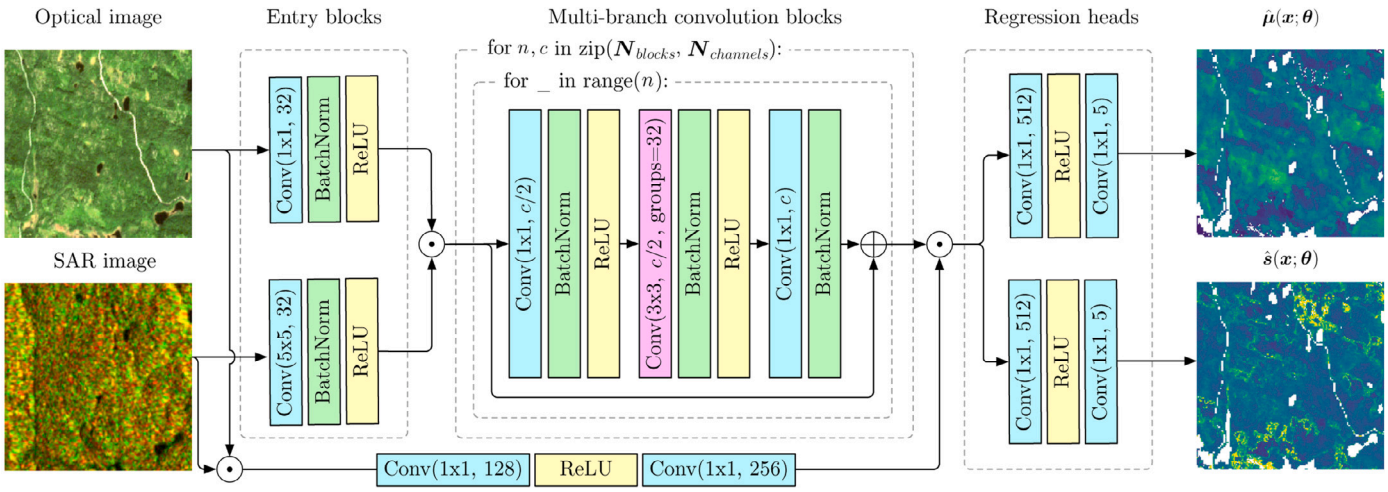

Fig. 5. Proposed model architecture. Convolutional layers are shown in cyan, with the kernel size and number of output channels in brackets. Batch normalization layers are in lime, ReLU activations in yellow and grouped convolutions in purple, with an additional argument for the number of groups. ⊕ and ⊙ denote element-wise addition, respectively concatenation along channel dimension, branching arrows indicate sharing ("copying") of the respective tensor. In our experiments we set $\boldsymbol{N}_{\text{blocks}} = [2, 3, 5, 3]$ and $\boldsymbol{N}_{\text{channels}} = [256, 512, 1024, 2048]$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

each other. Jointly, both satellites in each mission achieve high revisit frequencies of >1 visit every five days, depending on latitude. The Sentinel-2 satellites are equipped with multispectral instruments that detect light in 13 spectral bands ranging from visible blue to short-wave infrared, with band-dependent spatial resolutions between 10 and 60 m. This spectral profile gives rise to a multitude of applications, including forest and vegetation monitoring as well as the inference of vegetation parameters such as leaf area index or carbon stock. For this work, we have collected Sentinel-2 Bottom-of-Atmosphere (BOA) reflectance images (level 2 A, Main-Knorn et al., 2017) that have already been atmospherically corrected. For each ALS acquisition area, we select largely cloud-free images captured between May and October of the respective year, such that the area of interest is fully covered. Because BOA reflectance images still contain small amounts of atmospheric variation, we collect between two and seven images for each ground point (depending on cloud conditions), such that our model learns to cover a range of conditions.

The Sentinel-1 satellites are equipped with a C-band SAR sensor that actively monitors the surface with radiation of about 6 cm wavelength (European Space Agency, 2021b). Sentinel-1 is invariant against meteorological factors like clouds, and also against illumination conditions. Although C-Band SAR does not enter deep into the canopy, it may superficially penetrate it. Thus, we explore this as a complementary signal to the optical image. For every collected optical image, we query a Sentinel-1 image that was acquired within ten days of the optical image and that covers the same geographic area. As preprocessing, we carry out orbit correction and terrain correction, where for the latter we rely on the Copernicus digital elevation model with a resolution of one arc second. We intentionally keep the preprocessing simple because we expect our model to automatically infer the optimal transformations given the task at hand. For all preprocessing steps, we use SNAP (European Space Agency, 2020a) with the Sentinel-1 Toolbox (European Space Agency, 2020b) as provided by ESA.

4. Method

4.1. Forest structure model architecture

We utilize a multi-branch convolutional network (CNN) to transform co-registered optical and SAR images into maps of forest structure variables, and of their associated uncertainties. A CNN gradually transforms its inputs into outputs with a series of simple transformations (layers). One can think of the early layers as a pre-processing of the two input modalities, the middle layers as a feature extractor and the last layers as a regression, but there is no clear distinction between those parts. At the heart of all of them is the (discrete) convolution operator, i.e., a linear filtering of a c-channel input image with a $k \times k \times c$ kernel of weights. In each convolutional layer, multiple convolutions with different filter weights are applied, so the output is again an "image" with as many channels as there were filters. The filter weights are the trainable parameters of the model. Convolutional layers are interleaved with some element-wise, non-linear activation functions. The series of transformations gradually abstracts the bare image pixels into a sequence of feature maps until, at the output level, they have become the desired forest structure maps.

Our model architecture is inspired by the design principles of ResNeXt (Xie et al., 2017), modified to address the specific challenges of forest structure mapping. As in a number of remote sensing works (e.g., Rodríguez and Wegner, 2018; Lang et al., 2019), we retain the spatial dimensions of the input images throughout the network, i.e., we do not use any pooling operations or convolutions with stride. This is motivated by the observation that spatial feature aggregation tends to negatively impact performance on remote sensing images, as their spatial resolution is already limited and one cannot afford to lose further spatial detail. Furthermore, we have added separate *entry blocks* for the optical and SAR channels, allowing for domain-specific low-level processing. Finally, we have added a *global pixel shortcut* that re-injects the raw pixel values before the final regression layers, in order to better preserve high-frequency details that otherwise tend to get blurred (as information over an increasingly larger receptive field is mixed through repeated convolutions). We visualize our model architecture in Fig. 5 and in the following, the data flow in the model is described in more detail.

Entry blocks. Each entry block consists of a convolutional layer, followed by batch normalization (Ioffe and Szegedy, 2015) and a rectified linear unit (ReLU) non-linearity, which truncates negative values to zero. The convolution in the optical entry block has kernel size 1, i.e., it only mixes the channels at every individual pixel, while a larger kernel size of five is used in the SAR entry block to allow for learned spatial smoothing of the more noisy SAR data. For all convolution operations in our model that have kernel size >1, we apply appropriate zero-padding to preserve the spatial dimensions of the input.

Multi-branch convolution blocks. The outputs of both entry blocks are concatenated along their channel dimension and fed through N_{stages} *stages*, where each stage $i \in \{1 \dots N_{\text{stages}}\}$ consists of $N_{\text{blocks},i}$ ResNeXt *blocks*. Each ResNeXt block in turn is made up by a 1×1 convolution layer, a grouped convolution layer (Krizhevsky et al., 2012; Xie et al., 2017) with N_{groups} groups and kernel size three, and another 1×1 convolution layer. The former two are each followed by a batch normalization and ReLU, whereas the latter is followed only by batch normalization. Following the ResNet principle, a skip connection bypasses each block, such that the block learns an additive residual update to its input, making it possible to training much deeper networks (He et al., 2016). Due to the grouped convolution, each block effectively implements a multi-branch computational graph, where each branch can be interpreted as a lower-dimensional feature embedding, and where all branches are eventually combined by summation. This layout has been shown to offer superior predictive performance (Xie et al., 2017), while being less complex than the classical ResNet design (He et al., 2016).

Regression heads. Finally, from the resulting feature map, two parallel *heads* regress the parameters of the likelihood function for each structural variable and pixel. We assume a Gaussian likelihood, therefore for every pixel five means and log-variances are regressed. The variance of each predictive distribution will be trained to resemble aleatoric uncertainty, and is output in its logarithmic form for numerical stability (see Section 4.3). Each head consists of two convolution layers with kernel size one and an intermediate ReLU operation, gradually reducing the number of channels in the representation from 2304 down to five. In order to constrain the predicted mean values to their valid ranges (see Section 3), we apply a final $\exp(\cdot)$ activation to the `P95` and `MeanH` predictions, as well as a sigmoid activation $\sigma(x) = 1 \,/\, (1 + e^{-x})$ to the `Dens`, `Gini` and `Cover` means.

4.2. Model training

As commonly done in deep learning, we iteratively learn the model parameters with stochastic gradient descend, starting from a random initialization. In each iteration, we randomly sample a batch of $B = 64$ reference data patches of size 15×15 pixels, where a patch is only considered for training if the center pixel is *forested*. We consider a pixel forested if and only if it contains vegetation points (points with $Dz > 1.3\,\text{m}$, see Section 3.1) and also is considered forested based on NIBIO's Norway-wide timber volume map (Astrup et al., 2019). We use the latter as an additional precautionary measure to avoid unnecessary noise from non-forested areas, as we are interested in learning forest characteristics only. For every reference data patch, we randomly pick an optical image from the correct year and two SAR images (one ascending and one descending orbit) with acquisition dates near the one of the optical image. Using SAR with both ascending and descending orbits is expected to add robustness against terrain-induced geometric distortions such as shadowing, foreshortening and layover (Small et al., 1995; Carrasco et al., 1997). In total the model input is composed of 12 optical bands forming a tensor of size $B \times 12 \times 15 \times 15$, and four SAR bands (two polarizations × two orbital directions), forming a corresponding $B \times 4 \times 15 \times 15$ tensor. The model output are two tensors of shape $B \times 5 \times 15 \times 15$ each, one for the means of the five structural forest variables and one for their variances. An appropriate *loss function* measures the deviation between the model output and the ALS reference data, see below. Note that we only calculate this loss for pixels that are forested according to the above definition. Then, the gradients of all model parameters $\boldsymbol{\theta}$ w.r.t. that loss (plus some regularization term) are computed with back-propagation (Werbos, 1982) and used to update the parameters in the direction of steepest descent. We use the Adam (Kingma and Ba, 2015) variant of stochastic gradient descent (SGD), which adaptively scales the magnitude of the parameter updates based on the statistics of previous updates to speed up convergence. During training, we periodically evaluate the prediction error of the model (i.e., the current set of parameters) on a held-out validation set and keep the configuration $\boldsymbol{\theta}^*$ with the lowest error as the final model.

4.3. Loss function

The loss function, which is optimized during training, measures the quality of a set of network parameters $\boldsymbol{\theta}$ w.r.t. the training data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$, under some regularizing prior assumptions. We use a standard loss function $\mathcal{L}(\mathcal{D}; \boldsymbol{\theta})$ whose minimization corresponds to maximizing the posterior probability of the parameters given the training data. As it is commonly done in machine learning (see e.g. Goodfellow et al., 2016), we assume a zero-mean isotropic Gaussian prior over the network parameters (corresponding to $\mathcal{L}_2$ regularization) and a Gaussian likelihood function with mean $\hat{\boldsymbol{\mu}}_i := \hat{\boldsymbol{\mu}}(\boldsymbol{x}_i; \boldsymbol{\theta}) \in \mathbb{R}^5$ and diagonal covariance matrix with logarithmic elements $\hat{\boldsymbol{s}}_i := \hat{\boldsymbol{s}}(\boldsymbol{x}_i; \boldsymbol{\theta}) \in \mathbb{R}^5$:

$$\mathcal{L}(\mathcal{D}; \boldsymbol{\theta}) = \lambda \, \|\boldsymbol{\theta}\|_2^2 + \sum_{i,j} \left[\hat{s}_{ij} + \exp(-\hat{s}_{ij}) (\hat{\mu}_{ij} - y_{ij})^2 \right] \tag{1}$$

Here, $i \in \{1 \dots N\}$ indexes the data point while $j \in \{1 \dots 5\}$ indexes one of the five structure variables we are aiming to predict. The hyperparameter λ is inversely proportional to the variance of the parameter prior. Note that by explicitly predicting aleatoric uncertainty, in the form of log-variances $\hat{\boldsymbol{s}}(\boldsymbol{x}_i; \boldsymbol{\theta})$ of the output variables, the model learns to reduce the influence of data points on the loss that it deems particularly noisy, which in turn improves model performance (Kendall and Gal, 2017). Learning log-variances constrains the variances to positive values and prevents a potential division by zero in the loss function. The derivation and some further explanation of the loss function Eq. (1) is provided in Appendix A.1.

4.4. Acquiring uncertainty estimates

Let $\hat{\boldsymbol{\mu}}(\boldsymbol{x}_*; \boldsymbol{\theta})$ and $\hat{\boldsymbol{\sigma}}^2(\boldsymbol{x}_*; \boldsymbol{\theta})$ be the mean and variance of the predicted distribution that the network with parameters $\boldsymbol{\theta}$ outputs when shown the test image $\boldsymbol{x}_*$. Furthermore, let $p(\boldsymbol{\theta} \mid \mathcal{D})$ denote the posterior distribution over the network parameters, given the prior and the likelihood function defined in Section 4.3.

The exact predictive distribution $p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, \mathcal{D})$ of our model is intractable to compute, mainly due to the extremely high dimensionality of the parameter space. However, it is possible to *sample* from an approximate $p(\boldsymbol{\theta} \mid \mathcal{D})$ by training an ensemble of multiple neural networks from the same data (but with different random initializations and random batches for SGD). With M different networks this gives rise to a Monte Carlo approximation of the posterior,

$$p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, \mathcal{D}) \approx \frac{1}{M} \sum_{k=1}^{M} \mathcal{N}\left(\boldsymbol{y}_* \mid \hat{\boldsymbol{\mu}}_{*,k}, \operatorname{diag}(\hat{\boldsymbol{\sigma}}_{*,k}^2)\right), \tag{2}$$

where $\hat{\boldsymbol{\mu}}_{*,k} := \hat{\boldsymbol{\mu}}(\boldsymbol{x}_*; \boldsymbol{\theta}_k)$ and $\hat{\boldsymbol{\sigma}}_{*,k}^2 := \hat{\boldsymbol{\sigma}}^2(\boldsymbol{x}_*; \boldsymbol{\theta}_k)$ are the mean and variance predicted by the kth network.

In the approximate predictive distribution derived in Eq. (2), epistemic (i.e., model) uncertainty is captured by sampling multiple models $\boldsymbol{\theta}_k$, which will lead to widely scattered predictions and thus to high uncertainty in regions of the input space that are not sufficiently backed by training data. Aleatoric uncertainty, on the other hand, is learned from data and thus predicted as the variance of the likelihood function for each output variable at each image pixel. Note that for conceptual and computational simplicity, the likelihood is limited to have a diagonal covariance matrix. While this restricts the network's flexibility to express correlations in aleatoric uncertainty, we empirically find this modeling decision to be sufficient for the task at hand. Finally, based on Eq. (2), the tower rule and the law of total variance deliver approximations for the total mean and variance of the predictive distribution:

$$\mathbb{E}[\boldsymbol{y}_* \mid \boldsymbol{x}_*, \mathcal{D}] \approx \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\mu}}_{*,k} =: \bar{\boldsymbol{\mu}}_* \tag{3}$$

$$\operatorname{Var}[\boldsymbol{y}_* \mid \boldsymbol{x}_*, \mathcal{D}] \approx \frac{1}{M} \sum_{k=1}^{M} [\hat{\boldsymbol{\sigma}}_{*,k}^2 + (\hat{\boldsymbol{\mu}}_{*,k} - \bar{\boldsymbol{\mu}}_*)^2] =: \bar{\boldsymbol{\sigma}}_*^2 \tag{4}$$

In Section 5.2, we will empirically study the quality of the uncertainty estimate (4), and demonstrate how it can be used to identify and filter out unreliable predictions.

4.5. Implementation details

We have implemented our models in PyTorch (Paszke et al., 2017). We trained $M = 5$ models with batch size $B = 64$ and a base learning rate $\alpha = 10^{-4}$. The learning rate is automatically reduced by a factor of 0.1 when the validation loss has not improved for 15 consecutive epochs. We apply weight decay to control the strength of the unit Gaussian prior, with an empirically chosen magnitude of 10^{-3} that is inversely proportional to the hyperparameter λ from Eq. (1). We chose $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ as hyper-parameters for the Adam optimizer. Each neural network was trained on a single Nvidia RTX2080Ti GPU for ~14 days.

5. Experimental results and discussion

During evaluation, we slide a window of size 15×15 pixels over the test regions of our data. We use a stride of nine and only retain the innermost 11×11 pixels for every window location, thus allowing for a sufficiently large spatial input context for all predictions. Fig. 6 depicts the model output for a diverse 180×180 pixel example region sampled from the *East* area of the test set, along with the ALS reference data and the absolute error between the predicted mean and the reference data. Qualitatively, a high correspondence between the predicted mean and the reference data can be observed, indicating that the model has successfully learned to predict the respective variables for a very diverse sample region. Further, a correlation between predicted uncertainty and absolute error can be observed, although the latter is more "grainy", an attribute that we believe is transferred from the reference data, that can vary strongly between adjacent pixels (this is not that visible as in the error maps, because the overall value range is larger). The correlation between those quantities is not as visually obvious as the previous one, which calls for a more extensive investigation into the predicted uncertainties. Similar to Fig. 6, we have added qualitative examples from the *North* and *West* areas in Figs. B.12 and B.13 in the Appendix.

In the following subsections, among other experiments, we will quantitatively validate the observations made above (good reconstruction of the true structural variables as well as uncertainty estimates being representative of the error) for the entire test set.

5.1. Evaluation of forest variables prediction

For each variable, indexed by $j \in \{1 \dots 5\}$, we report the Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the Mean Bias Error (MBE) defined as follows:

$$\mathrm{MAE} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} |\bar{\mu}_{ij} - y_{ij}| \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (\bar{\mu}_{ij} - y_{ij})^2} \tag{6}$$

$$\mathrm{MBE} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \bar{\mu}_{ij} - y_{ij} \tag{7}$$

To represent the predictive distribution for each pixel i and variable j as a point estimate, we use its approximate mean $\bar{\mu}_{ij}$ in accordance with Eq. (3). The MAE and RMSE metrics measure the average deviation between model prediction and reference data (with RMSE giving higher penalty to large deviations), while the MBE serves to identify systematic biases in the predictions. In addition to the above metrics, we calculate their *normalized* counterparts MAE%, RMSE% and MBE% by dividing by the corresponding mean values over the training data.

Table 1 shows the results of the test set evaluation of the final model. Overall (subtable a), we achieve low mean absolute errors for all predicted variables, e.g., 1.65 m for the 95th height percentile (P95). Assessing the relative errors, the experiments show that the developed model produces consistently accurate predictions with an MAE% always smaller than 15%. The variables ranking in order of increasing relative error (MAE%) are forest cover (`Cover`, 11.2%), vegetation density (`Dens`, 11.8%), Gini index (`Gini`, 12.4%), the 95th height percentile (P95, 12.9%) and lastly the mean height (`MeanH`, 14.4%). When evaluating the individual geographic regions shown in Fig. 2, we observe comparable values for the eastern region (subtable b), which can be explained by this region making up for the majority of reference data points. Interestingly, the model performs even better in the northern test region in terms of mean absolute error of P95 and `MeanH`, however the error for the remaining structural variables is slightly bigger. For the western test region, evaluation errors are consistently higher, yielding e.g. an MAE in P95 of 1.85 meters and an MAE% of 14.4%. Possible reasons for the slightly lower performance in the western region are the scarcity of cloud-free data due to oceanic climate, and the increased topographic complexity. The fjord landscape in these western areas is characterized by very steep and narrow valleys causing both topographic shading and large variations in sun-target-sensor geometry. All of these factors have a negative impact on the quality of the satellite observations and thus the predictions obtainable in such areas. In addition, the forests in these areas are more diverse in terms of species compositions and structurally complex compared to the other areas. It is noteworthy that, for all of the tested regions, the MBE% is <0.7%, indicating the absence of any systematic model bias, regardless of the predicted variable and geographic region. Our model can therefore be considered an unbiased estimator for forest structure, a significant benefit for downstream tasks such as decision making based on the resulting maps.

For further analysis of our model, we provide confusion plots for all variables in Fig. 8. For the plots, all test samples are binned with regard to both their true, ALS-derived reference value (x axis) and their predicted mean (y axis), the color of a bin indicating how many samples fall into the respective bin. Higher densities along the identity line imply higher accuracy of the model, and indeed we can observe very high agreement between prediction and reference data, for all variables.

In addition, we conduct a residual structure analysis of each variable with respect to all others, i.e., we investigate how the residuals in one variable are distributed across other variables' value ranges. Note that to get more stable estimates, we remove data points exceeding the 99th percentile of the respective variable before binning. For this analysis, we used $b = 10$ bins. In Fig. 7, the column indicates the *query* variable, whose ground truth values (clipped to the 99th percentile for stability) are discretized into b bins and define the ordering along the x-axis. The row indicates the *target* variable whose residual distribution is plotted (so, for instance, the fourth graph in the top row are the P95 residuals ordered by the corresponding pixels' `Gini` coefficient). The solid green line denotes the mean residual, the shaded area the residuals' standard deviation. Mean residuals close to zero (gray line) mean that the estimation of the target is unbiased at the given value of the query. We observe that our predictions generally have low bias across a large portion of the range, which is consistent with our findings in Section 5.1. A notable exceptions is the under-estimation of the height-related variables P95 and `MeanH` on high trees – a well-known effect when retrieving vegetation height from satellite images (Potapov et al., 2021; Lang et al., 2019). A similarly pronounced under-estimation bias at the top of their own value range can be noticed for the canopy density (`Dens`) and for the Gini coefficient (`Gini`); moreover, high `Gini` values also tend to cause underestimation of P95. In the opposite direction, we also observe some (weaker) over-estimation biases for low reference values. This concerns in particular the `Dens` and `Cover` estimates at low values of `Dens`, `Gini` and `Cover`.

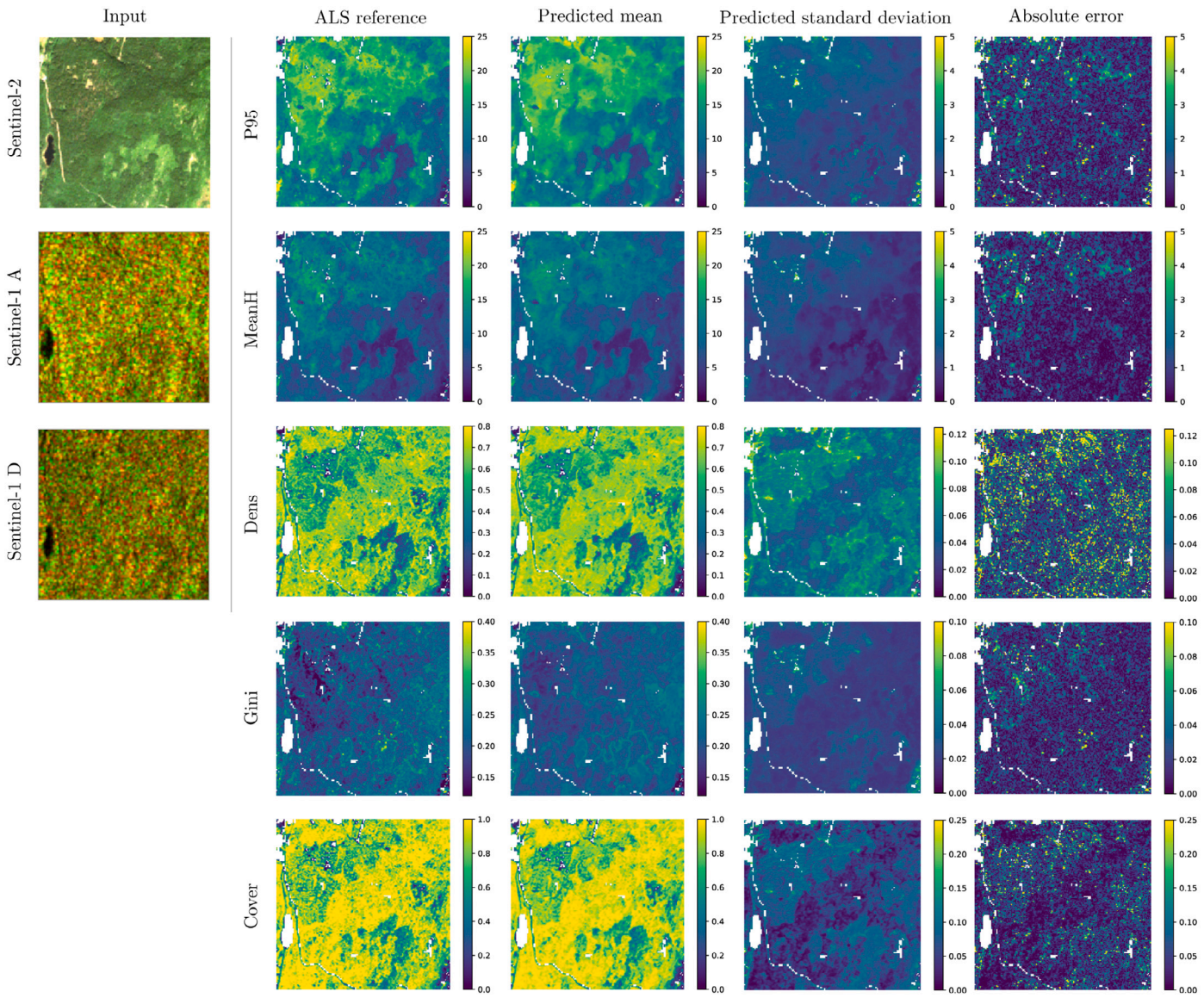

Fig. 6. Example result from the test set of the *East* area (180 × 180 pixels, 324 ha). *Left tile*: The model input is composed of a Sentinel-2 optical image (12 bands), and two preprocessed Sentinel-1 SAR images taken from an ascending and descending orbit direction, respectively. For visualization, the amplitudes of the VH and VV polarizations have been assigned to the red and green channels of the RGB image. *Right tile*: ALS reference data, the predicted mean and standard variation as well as the absolute prediction error $|\bar{\mu}_{ij} - y_{ij}|$ for each of the five structural variables. The figure shows high correspondence between reference data and the predicted mean, indicating that our model is able to accurately regress the structural variables. There is also a clear (albeit not as crisp) correlation between predicted standard variation and absolute error, indicating that erroneous predictions are assigned higher uncertainty. We analyze the latter in more detail in Section 5.2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Overall, the biases are moderate and exhibit the typical behavior of regressors fitted to minimize mean squared error. That is, they tend towards the data mean and, when the evidence is weak or ambiguous, over-estimate low values and under-estimate high ones.

5.2. Uncertainty evaluation

We evaluate the quality of the model's uncertainty estimates (Eq. (4)), i.e., how well the predicted uncertainty correlates with the expected error of a given prediction. If a model is perfectly calibrated, then, for any value of the predicted variance, the expected squared error between the true value and the predicted mean should be equal to the predicted variance (by definition of the variance). To obtain an estimate of the expected squared error, we group predictions into bins based on the predicted uncertainty and compare, for each bin, the mean squared error with the mean predicted variance of all samples in the bin. For a more intuitive interpretation, in the units of the original variables, we take square roots and compare the empirical RMSE and the root mean variance. In Fig. 9, the result of this comparison is shown for 20 bins for the P95 variable. It can be observed that the plot corresponding to the ensemble follows the identity line closely, implying highly reliable uncertainty estimates. Only for higher-variance predictions ($\mathrm{Var}[\boldsymbol{y}_* \mid \boldsymbol{x}_*, D] \gtrsim 12$) we can observe a slight under-estimation of uncertainty. Each of the individual neural networks, in contrast to the ensemble, suffers from much more severe under-estimation of uncertainty (i.e., they are over-confident) throughout all uncertainty levels. This is consistent with the literature, where it has repeatedly been observed that neural network models tend to

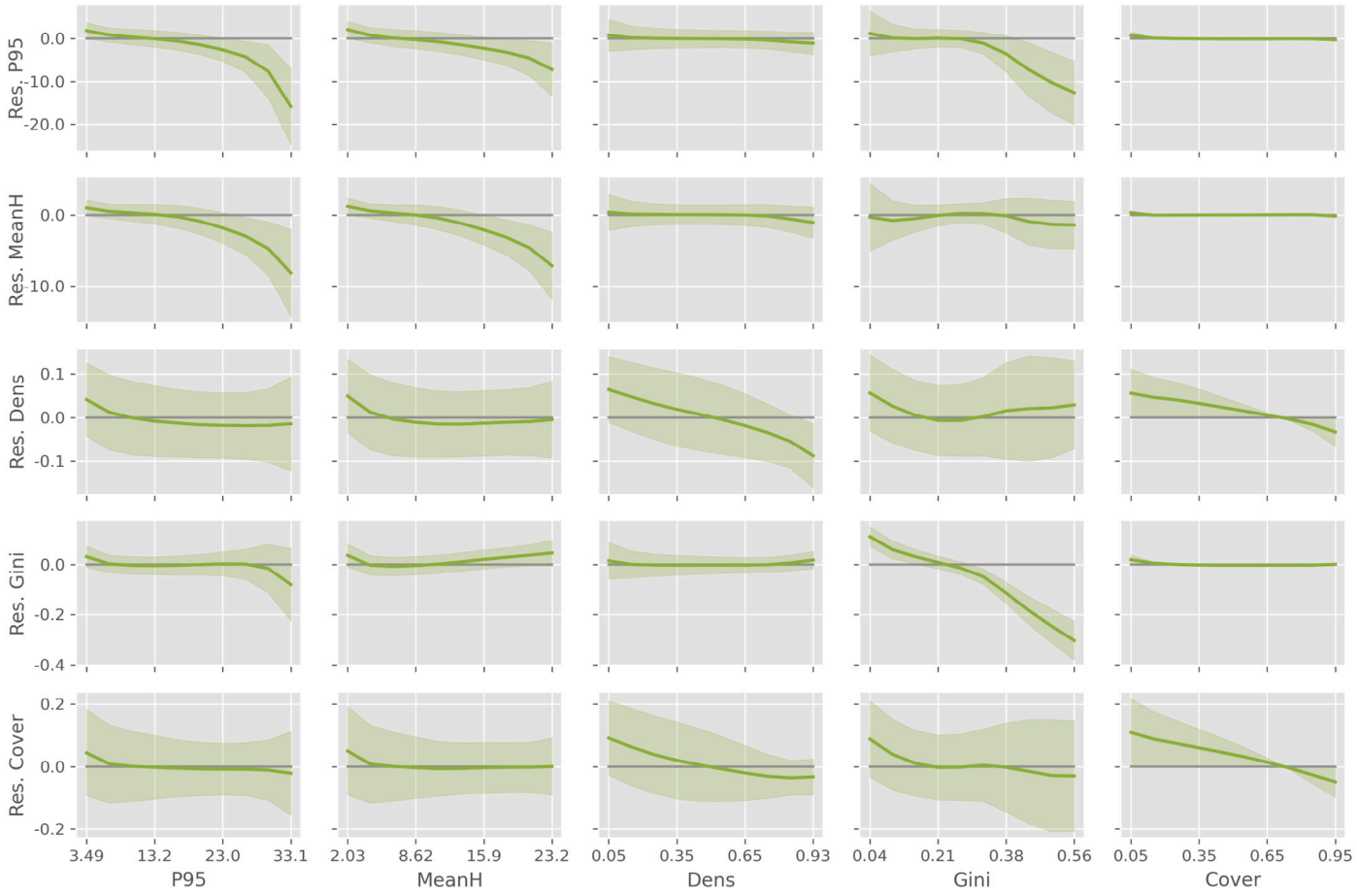

Fig. 7. Residual structure analysis for all variables. Columns indicate the variable we bin according to (*query*), and rows report the residuals of each variable in the respective bins (*target*). Additionally, in shaded we show the respective standard deviation of the residuals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

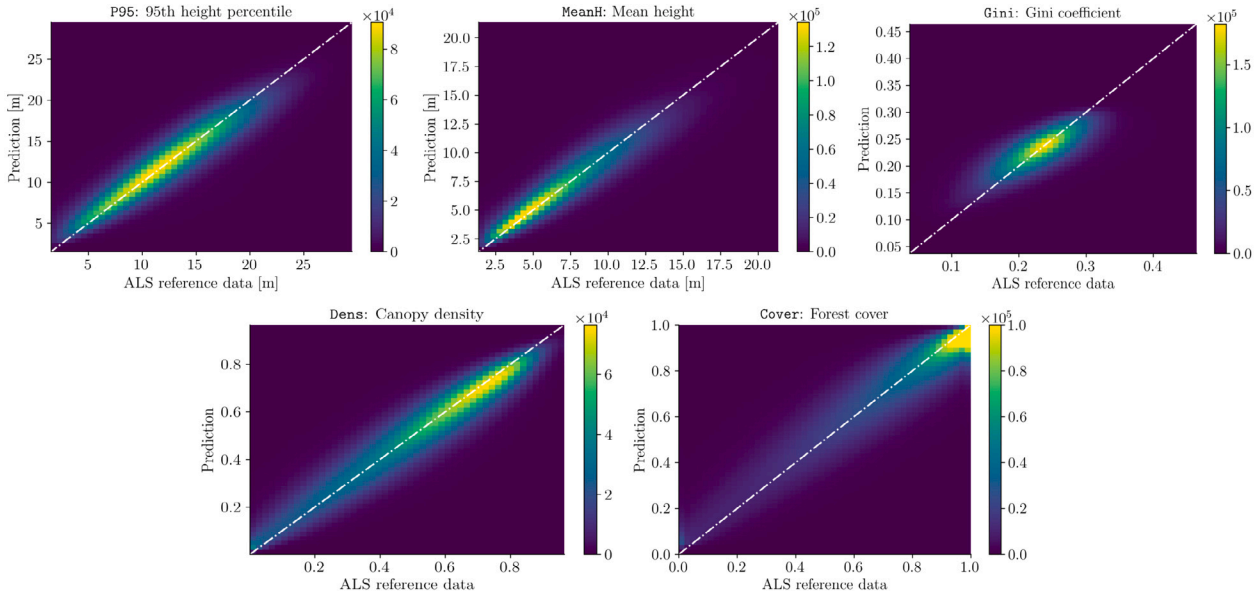

Fig. 8. Confusion plots for the predicted structure variables. In each plot, the *x* axis denotes the ALS reference data value and the *y* axis denotes the mean of the predicted likelihood distribution. The number of data points that fall into each 2D bin is indicated by the color. The plots demonstrate that most predictions are located very close to the identity line, where prediction and ALS reference data agree. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Test set evaluation results of the proposed model for all ALS areas (subtable a) as well as for the individual geographic regions shown in Fig. 2 (subtables b–d). P95 and MeanH metrics are given in meters, while Dens, Gini and Cover are reported as fractions. All normalized metrics (postfixed with %) are given as a fraction over the respective training set mean.

		P95	MeanH	Dens	Gini	Cover
(a) All regions	MAE	1.648	1.127	0.061	0.028	0.078
	MAE%	0.129	0.144	0.118	0.123	0.112
	RMSE	2.298	1.595	0.082	0.039	0.107
	RMSE%	0.179	0.204	0.158	0.170	0.154
	MBE	−0.086	−0.040	−0.003	−0.001	0.001
	MBE%	−0.007	−0.005	−0.006	−0.004	0.002
(b) East	MAE	1.631	1.127	0.060	0.027	0.075
	MAE%	0.127	0.144	0.114	0.121	0.107
	RMSE	2.250	1.589	0.080	0.038	0.101
	RMSE%	0.176	0.203	0.153	0.166	0.145
	MBE	−0.090	−0.040	−0.004	−0.001	0.000
	MBE%	−0.007	−0.005	−0.007	−0.005	0.000
(c) North	MAE	1.451	0.904	0.071	0.031	0.106
	MAE%	0.113	0.115	0.136	0.136	0.152
	RMSE	2.238	1.329	0.095	0.042	0.140
	RMSE%	0.175	0.170	0.182	0.186	0.200
	MBE	−0.068	−0.038	0.003	0.001	0.009
	MBE%	−0.005	−0.005	0.005	0.005	0.013
(d) West	MAE	1.845	1.224	0.069	0.031	0.091
	MAE%	0.144	0.156	0.132	0.136	0.131
	RMSE	2.609	1.740	0.093	0.042	0.127
	RMSE%	0.204	0.222	0.178	0.184	0.182
	MBE	−0.069	−0.041	−0.001	−0.000	0.003
	MBE%	−0.005	−0.005	−0.002	−0.001	0.005

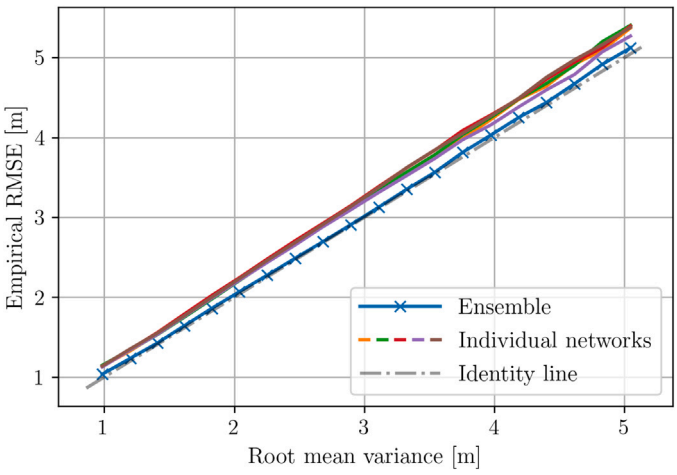

Fig. 9. Calibration plot for the P95 variable, using 20 uncertainty bins. For the ensemble, the predicted uncertainty agrees very well with the actually observed error, whereas we observe systematic over-confidence in the individual models' predictions. Note that for this plot, we have removed the first and last percentile of predictions (according to predicted uncertainty) to avoid bins with a very small number of samples at the tails of the distribution, because with very few samples the MSE is a noisy estimate of the expected squared error.

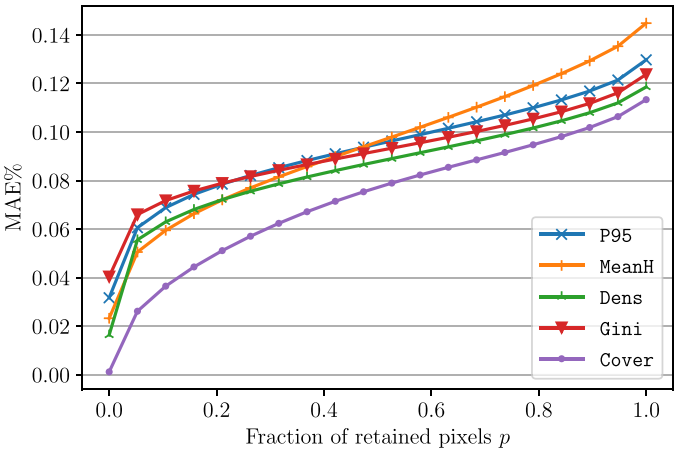

Fig. 10. MAE% for all variables measured on a fraction $p \in (0, 1]$ of test pixels that have the lowest associated uncertainty (as defined by Eq. (4)). The plot shows a strong correlation between estimated uncertainty and expected normalized absolute error. We show the normalized MAE% (instead of the absolute MAE) so we can compare the results for the five structural variables in a single plot.

be overconfident about their predictions (Guo et al., 2017; Lakshminarayanan et al., 2017). The calibration plots for the remaining structure variables are qualitatively similar and can be found in the Appendix (Fig. B.14). Note that even though the ensemble reaches the best calibration for all structure variables (compared to the individual neural networks), the individual networks' uncertainty estimates are, in our specific case, already very good. We attribute this to the large and diverse dataset that was available for training, such that only few test samples are insufficiently represented by the training set and require the (approximate) Bayesian marginalization to achieve good model calibration.

Another way to validate the usefulness of the predicted uncertainty values is to observe how error metrics change when retaining a decreasing fraction $p \in (0, 1]$ of least uncertain predictions. The intuition is that if one discards predictions with high uncertainty and the latter is indeed a good predictor of the expected error, then the average error of the remaining predictions should decrease. Fig. 10 displays this relationship between the fraction of retained pixels and the MAE%, for all structural variables. Indeed, all curves increase monotonically, implying that the predicted uncertainty correlates (in expectation) with the actual error at a given test sample.

5.3. Sensor ablation study

We investigate how much the reported accuracy depends on the individual optical and SAR image inputs. To this end, Table 2 reports the MAE when training a network with different input data configurations. Unsurprisingly, the default configuration that uses one Sentinel-2 optical and two Sentinel-1 SAR images ("S2+S1") performs best w.r.t. all predicted structure variables. What stands out is that removing the optical information ("S1 only") is significantly more detrimental than removing the SAR input ("S2 only"), indicating that the former is more predictive of forest structure. We assume that this is mainly due to two reasons: (1) Sentinel-1 imagery is noisier, which impairs the analysis of small structures at the level of individual pixels and below; and (2) Sentinel-1 is more heavily affected by topographic influences. In particular, the mountainous topography prevalent in Norway can give rise to SAR-specific effects, such as shadowing, foreshortening and layover (Small et al., 1995; Carrasco et al., 1997).

The ablation study further shows that including SAR data from both ascending and descending orbits is beneficial in terms of regression performance, compared to using only one (randomly chosen) orbit direction during training and testing.[1] As an explanation, we suspect

[1] We argue that choosing the orbit direction randomly is the most natural way of disregarding that factor, without introducing biases specific to a given orbit direction.

Table 2
MAE of the reported structural variables for various input configurations. The version that uses both Sentinel-2 optical and Sentinel-1 SAR imagery performs best, while leaving out the optical input is more harmful compared to leaving out the SAR. The experiments also demonstrate the effectiveness of using SAR images from both ascending and descending orbits, as compared to only one, random orbit direction ("S1Rand").

Input	P95	MeanH	Dens	Gini	Cover
S2+S1	**1.697**	**1.162**	**0.063**	**0.029**	**0.080**
S2+S1Rand	1.778	1.219	0.066	0.029	0.083
S2 only	1.813	1.243	0.067	0.029	0.084
S1 only	3.052	2.123	0.120	0.038	0.148
S1Rand only	3.480	2.421	0.144	0.040	0.174

a combination of two effects: (1) The model has a second observation for any given pixel, thus benefiting from redundancy to suppress noise; and (2) the scene is illuminated from two different directions, which helps to mitigate SAR-specific effects as outlined previously.

Lastly, the ablation study demonstrates that the impact of removing one of the SAR images is far less pronounced when additional optical information is available ("S2+S1" vs. "S2+S1Rand"), compared to a configuration that solely relies on SAR ("S1 only" vs. "S1Rand only"). This outcome further supports the finding that optical images are the more predictive data source for forest structure.

5.4. Country-wide forest structure map

To demonstrate the applicability of our method at country-scale, we compute a Norway-wide map of forest structure variables for the year 2020 and make it publicly available. To be more robust against clouds and to avoid gaps, we use $T = 10$ optical images per location that were acquired throughout the leaf-on-period June to October. We also add one SAR image to every location and pair it with the optical data. We feed all eleven inputs per location to our model, which consists of five independently trained neural networks as described in Section 4. This ultimately yields ten predictions for mean and variance for every geographical location. To produce an easy-to-interpret map, we collapse the ten predictive distributions into a single point estimate by applying an inverse-variance weighting:

$$y_{ij} = \frac{\sum_{t=1}^{T} \bar{\mu}_{tij}/\bar{\sigma}^2_{tij}}{\sum_{t=1}^{T} 1/\bar{\sigma}^2_{tij}}, \tag{8}$$

for every pixel i and structural variable j.

This scheme poses a natural way of combining multiple measurements into a single forest structure map, exploiting the (previously demonstrated) property of our model to return well-calibrated uncertainties. Intuitively, predictions that are assigned a high uncertainty by our model should have less influence on the final results compared to predictions of high confidence. It can be shown that the resulting, inverse-variance weighted forest structure estimates have the smallest expected error among all possible weighted averages (Strutz, 2010). Therefore, by means of outputting well-calibrated predictions, we are able to combine multiple measurements in an theoretically optimal way, which is not possible when only point estimates are available.

Fig. 11 shows overview images of the generated map for each of the five structural forest variables. We mask out areas that are not considered forested according to the NIBIO timber volume map (see Section 4.2) as already done for training our model.[2]

[2] Note that two small regions in northern Norway are missing in our map due to missing data in the forest mask.

6. Conclusion

The main finding of our research is that forest structure indicators can be mapped at country-scale from publicly available Sentinel data, in combination with modern deep machine learning and open airborne laser scanning (ALS) data as reference data for training. The availability of these maps at almost any point in time (starting from the beginning of the Copernicus program in 2015) opens up new possibilities for large scale forest monitoring and resource mapping and is thus a promising tool for understanding and tackling challenges such as climate change or biodiversity loss. A straightforward application of the predicted structural variables could be updating existing nationwide forest resource maps hitherto produced on the basis of the same variables, but derived from ALS data (Nord-Larsen and Schumacher, 2012; Nilsson et al., 2017; Astrup et al., 2019). Given the large costs for updating these maps, our Sentinel-derived forest structure variables could provide inexpensive auxiliary data for producing dense time series of forest resource maps. The availability of such dynamic maps can help to better understand the effect of forest management practices on forest biomass dynamics and on the variations in functional diversity. In addition to their use for mapping traditional variables (e.g., above-ground biomass), dense time series of canopy height (e.g., P95) from Sentinel data allow one to derive the canopy height vertical growth, which in turn can be used to measure site productivity (Lennart Noordermeer and Næsset, 2018; Svein Solberg and Puliti, 2019). In addition, the produced maps can also be used to provide proxies to identify and map forest disturbances on a large scale (Hansen et al., 2013; Senf et al., 2018).

An innovative aspect of our method is that it additionally predicts variables describing the vegetation density (Dens) and its variation in the vertical profile of the canopy (Gini), different to previous studies that focused on canopy height or cover only. These additional variables are important as they provide complementary information to the height and the cover and thus allow for a more comprehensive understanding of forest structure on a large scale. The combined use all of these variables further opens up new possibilities to address complex issues, such as the quantification of biomass losses and gains from subtle land-use changes, e.g., forest degradation and forest restoration.

A further novelty of the presented method, compared to previous applications of deep learning to remotely sensed data, is that it outputs pixel-wise, well-calibrated output distributions instead of mere point estimates. The quantification of the predictive uncertainty makes the system more reliable and more trustworthy. In particular, estimates of the expected predictive error can be propagated to downstream tasks, such as higher-level mapping systems that rely on our forest structure estimates as input. Even further downstream, maps ultimately serve as a basis for forest management, where well-calibrated uncertainty estimates ("knowing what we don't know") translate to better decision making.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Derivations

In the following, we provide mathematical derivations for the employed loss function (Eq. (1)) and the moments of the posterior predictive distribution (Eqs. (3) and (4)).

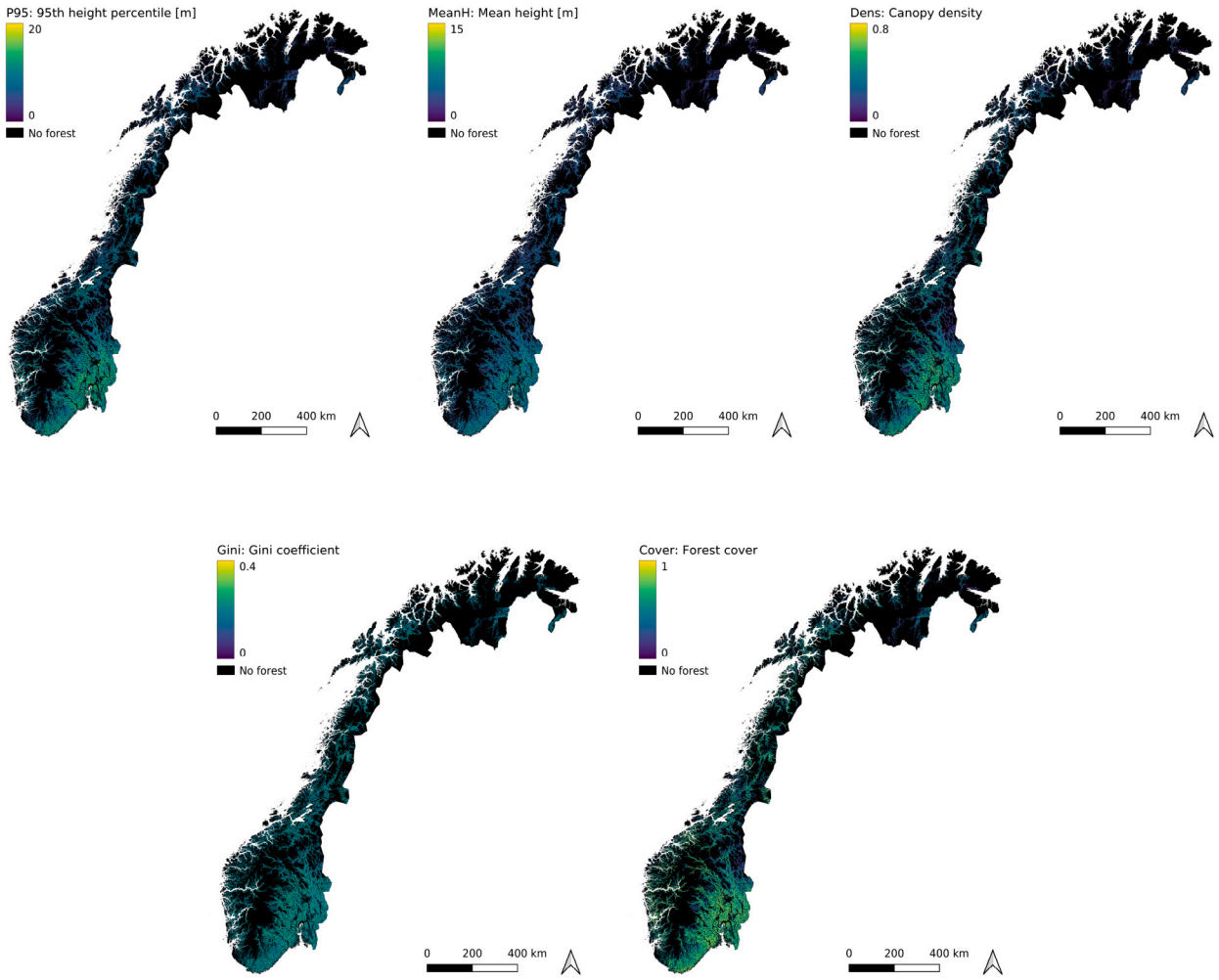

Fig. 11. Overview images of the five channels of the produced Norway-wide forest structure map, corresponding to the five structural variables. Non-forested areas have been masked out and labeled as "No forest".

A.1. Maximization of the parameter posterior

A standard approach in machine learning (see e.g. Goodfellow et al., 2016), we start with the maximizer of the logarithmic posterior parameter probability (i.e., its mode) and then deduce the equivalence to the loss function presented in Eq. (1):

$$\arg\max_{\theta} \log p(\theta \mid \mathcal{D}) \tag{A.1a}$$

$$= \arg\max_{\theta} \log \underbrace{p(\theta)}_{\text{prior}} + \log \underbrace{p(\mathcal{D} \mid \theta)}_{\text{likelihood}} - \log \underbrace{p(\mathcal{D})}_{\text{evidence}} \tag{A.1b}$$

$$= \arg\max_{\theta} \log p(\theta) + \log p(\mathcal{D} \mid \theta) \tag{A.1c}$$

$$= \arg\min_{\theta} - \log p(\theta) - \sum_{i=1}^{N} \log p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \theta) \tag{A.1d}$$

$$= \arg\min_{\theta} - \log \mathcal{N}(\theta; \mathbf{0}, \sigma_p^2 I) - \sum_{i=1}^{N} \log \mathcal{N}(\boldsymbol{y}_i; \hat{\boldsymbol{\mu}}_i, \hat{\Sigma}_i) \tag{A.1e}$$

$$= \arg\min_{\theta} \frac{\lambda}{2} \|\theta\|_2^2 + \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\Sigma}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i) \tag{A.1f}$$

$$= \arg\min_{\theta} \underbrace{\lambda \|\theta\|_2^2 + \sum_{i,j} \left[\hat{s}_{ij} + \exp(-\hat{s}_{ij})(\hat{\mu}_{ij} - y_{ij})^2\right]}_{\mathcal{L}(\mathcal{D};\theta)}. \tag{A.1g}$$

In Eqs. (A.1b) and (A.1c), we apply Bayes' rule (in log space) and then drop the evidence term as it does not depend on θ. We assume independent and identically distributed labels given the respective input and the model parameters and can thus factor the likelihood into individual data point likelihoods (Eq. (A.1d)). In Eq. (A.1e), $\mathcal{N}(\cdot \mid \boldsymbol{\mu}, \Sigma)$ denotes the probability density function of the multivariate Gaussian normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ. We assume an isotropic Gaussian with variance σ_p^2 as prior over the parameters, as it is common in machine learning (this is equivalent to $\mathcal{L}_2$ regularization). In the same line, $\hat{\Sigma}_i = \mathrm{diag}(\hat{\boldsymbol{\sigma}}_i^2)$ is a diagonal matrix containing the predicted variances for all structural variables as diagonal elements. To arrive at Eq. (A.1f), we plug in the Gaussian density function and observe that $\lambda \propto 1/\sigma_p^2$ governs the strength of regularization. In Eq. (A.1g), $j \in \{1 \dots 5\}$ indexes the structure variables and we again use $\hat{s}_{ij} = \log \hat{\sigma}_{ij}^2$.

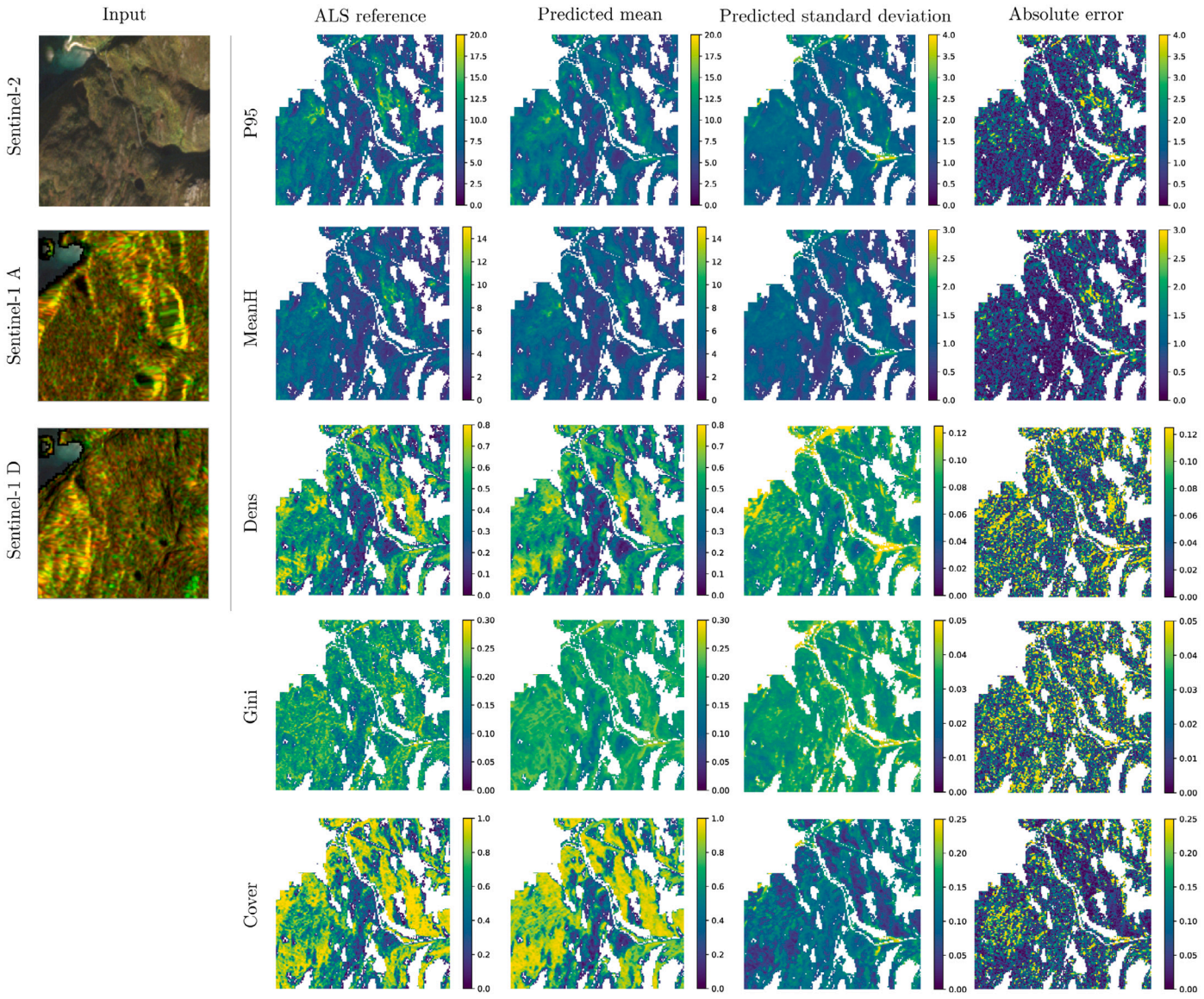

Fig. B.12. Qualitative prediction example (180 × 180 pixels, 324 ha) from the *North* region, analogous to Fig. 6.

A.2. Approximate Bayesian marginalization

The posterior predictive distribution $p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, \mathcal{D})$ is obtained by marginalizing over the posterior parameter distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$:

$$p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, \mathcal{D}) \tag{A.2a}$$

$$= \int p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{\theta} \tag{A.2b}$$

$$= \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathcal{D})} \left[p(\boldsymbol{y}_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}) \right] \tag{A.2c}$$

$$\approx \frac{1}{M} \sum_{k=1}^{M} \mathcal{N}\left(\boldsymbol{y}_* \mid \hat{\boldsymbol{\mu}}_{*,k}, \operatorname{diag}(\hat{\boldsymbol{\sigma}}^2_{*,k})\right), \tag{A.2d}$$

where $\hat{\boldsymbol{\mu}}_{*,k} := \hat{\boldsymbol{\mu}}(\boldsymbol{x}_*; \boldsymbol{\theta}_k)$ and $\hat{\boldsymbol{\sigma}}^2_{*,k} := \hat{\boldsymbol{\sigma}}^2(\boldsymbol{x}_*; \boldsymbol{\theta}_k)$ and $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta} \mid \mathcal{D})$. Due to the intractability of the integral in Eq. (A.2b), we estimate this quantity using Monte Carlo sampling (Eq. (A.2d)). In practice, we sample multiple $\boldsymbol{\theta}_k$ by training an ensemble and treating each neural network as a sample from an approximate $p(\boldsymbol{\theta} \mid \mathcal{D})$ (Gustafsson et al., 2020; Wilson and Izmailov, 2020). Given the resulting approximate predictive distribution (Eq. (A.2d)), which is a mixture of Gaussians, the mean is obtained by the tower rule,

$$\mathbb{E}[\boldsymbol{y}_* \mid \boldsymbol{x}_*, \mathcal{D}] = \mathop{\mathbb{E}}_{\boldsymbol{\theta}|\mathcal{D}} \Big[\mathbb{E}[\boldsymbol{y}_* \mid \boldsymbol{\theta}, \boldsymbol{x}_*] \Big] \approx \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\mu}}_{*,k}, \tag{A.3}$$

and the variance by the law of total variance:

$$\operatorname{Var}[\boldsymbol{y}_* \mid \boldsymbol{x}_*, \mathcal{D}] \tag{A.4a}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{\theta}|\mathcal{D}} \Big[\operatorname{Var}[\boldsymbol{y}_* \mid \boldsymbol{\theta}, \boldsymbol{x}_*] \Big] + \mathop{\operatorname{Var}}_{\boldsymbol{\theta}|\mathcal{D}} \Big[\mathbb{E}[\boldsymbol{y}_* \mid \boldsymbol{\theta}, \boldsymbol{x}_*] \Big] \tag{A.4b}$$

$$\approx \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\sigma}}^2_{*,k} + \frac{1}{M} \sum_{k=1}^{M} (\hat{\boldsymbol{\mu}}_{*,k} - \bar{\boldsymbol{\mu}}_*)^2 \tag{A.4c}$$

$$= \frac{1}{M} \sum_{k=1}^{M} [\hat{\boldsymbol{\sigma}}^2_{*,k} + (\hat{\boldsymbol{\mu}}_{*,k} - \bar{\boldsymbol{\mu}}_*)^2], \tag{A.4d}$$

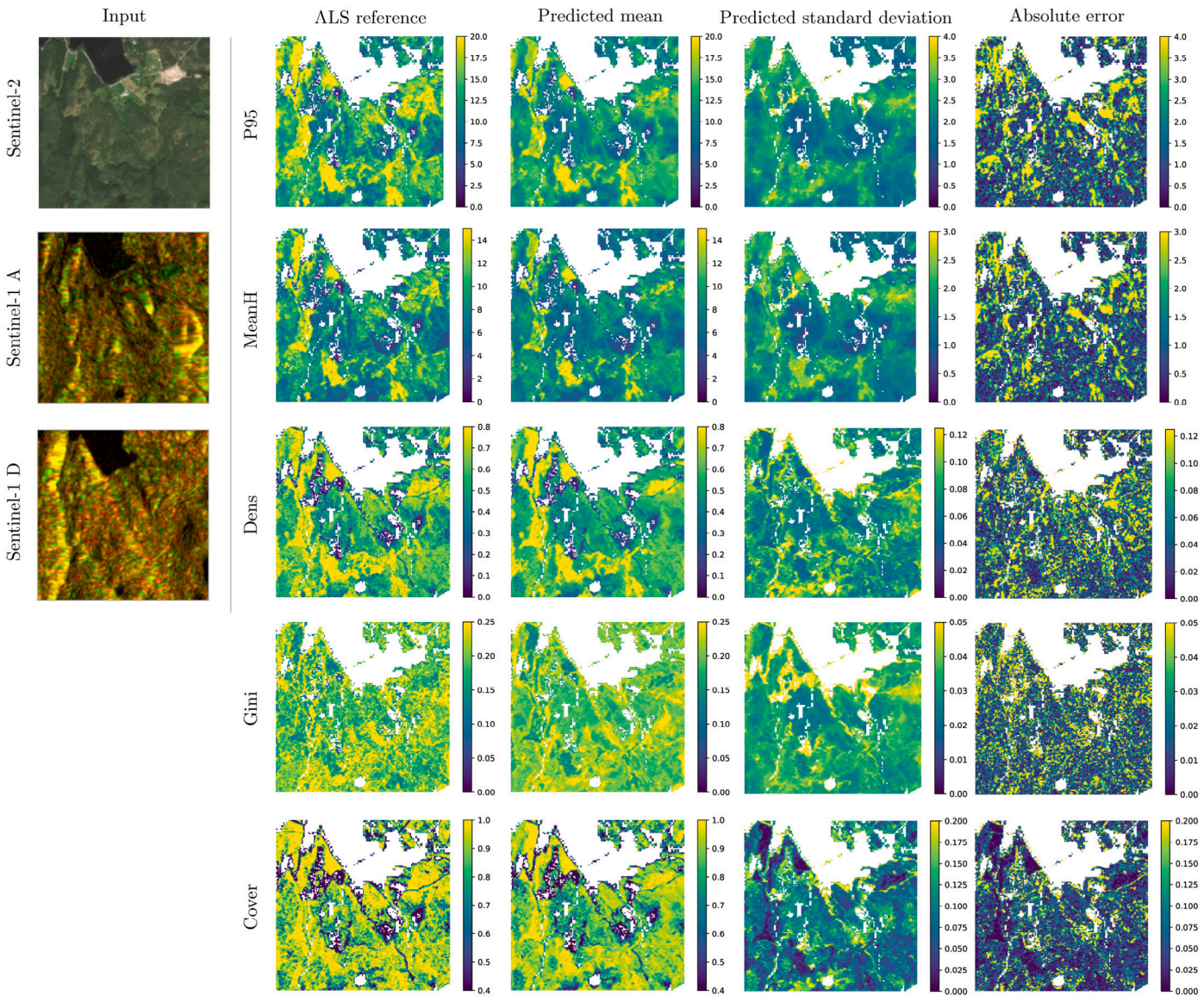

Fig. B.13. Qualitative prediction sample (180 × 180 pixels, 324 ha) from the *West* region, analogous to Fig. 6.

where we again abbreviate $\bar{\mu}_* := \sum_{k=1}^{M} \hat{\mu}_{*,k}$.

Appendix B. Additional experimental results

Further ablation study results. Supplementary to Table 2, where we reported the MAE of our method for various input configurations, Table B.3 show the other error metrics for these experiments.

Further calibration plots. Fig. B.14 shows the calibration plots similar to Fig. 9 for the remaining structure variables. As for P95, we observe a systematic under-estimation of the variance for the individual networks, and a clearly better uncertainty calibration for the ensemble. The effect is more or less pronounced, depending on the variable. For all plots we have again used 20 uncertainty bins and the same noise removal technique as explained for Fig. 9.

Baseline comparison. At the time of writing, we are not aware of any comparable method that predicts diverse forest structure variables from optical and SAR images, and that we could compare to. As a baseline to better understand how our method fares in comparison to a common "best practice" method for vegetation mapping, we train a random forest (RF) regressor on our data, using per-pixel spectral intensities as inputs. Because of the size of the training dataset (≈65 million data points), we resort to bootstrapping with a sample ratio of 0.05 and 20 ensemble members. Table B.4 compares the test set results of the RF regressor to those of our method. In terms of both MAE and RMSE, the proposed neural network outperforms the RF baseline by large margins for all five structure variables, in most cases reducing the deviations from ground truth by 40–50% (except for `Gini`, where the gains are only 25–30%). The MBEs are generally low, underlining that both methods are unbiased within the expectable measurement accuracy: all relative MBE-values lie within ±2%. Overall, the experiment confirms the superior predictive power of deep feature extractors, in particular when trained on large data sets.

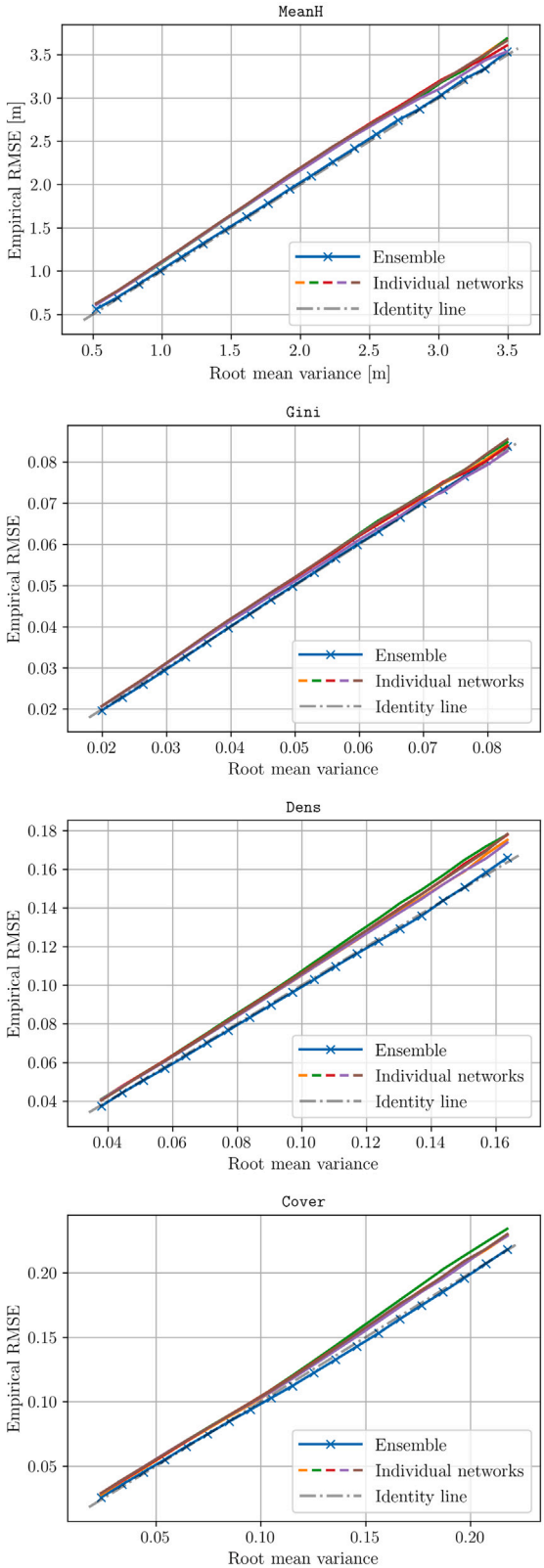

Fig. B.14. Calibration plots for the `MeanH`, `Gini`, `Dens` and `Cover` variables, similar to those in Fig. 9.

Table B.3
Detailed experimental results of the input ablation study conducted in Section 5.3, reporting the performance of our method on different input configurations.

		P95	MeanH	Dens	Gini	Cover
(a) S2+S1	MAE	1.697	1.162	0.063	0.029	0.080
	MAE%	0.132	0.148	0.121	0.125	0.115
	RMSE	2.366	1.645	0.085	0.039	0.110
	RMSE%	0.185	0.210	0.163	0.172	0.158
	MBE	−0.039	−0.019	−0.002	−0.000	−0.001
	MBE%	−0.003	−0.002	−0.004	−0.002	−0.001
(b) S2+S1Rand	MAE	1.778	1.219	0.066	0.029	0.083
	MAE%	0.139	0.156	0.126	0.128	0.119
	RMSE	2.470	1.723	0.089	0.040	0.114
	RMSE%	0.193	0.220	0.170	0.175	0.164
	MBE	−0.075	−0.041	−0.004	−0.001	0.002
	MBE%	−0.006	−0.005	−0.007	−0.003	0.003
(c) S2 only	MAE	1.813	1.243	0.067	0.029	0.084
	MAE%	0.141	0.159	0.129	0.129	0.121
	RMSE	2.511	1.753	0.091	0.040	0.116
	RMSE%	0.196	0.224	0.174	0.177	0.167
	MBE	−0.037	−0.006	−0.000	−0.001	0.005
	MBE%	−0.003	−0.001	−0.001	−0.003	0.007
(d) S1 only	MAE	3.052	2.123	0.120	0.038	0.148
	MAE%	0.238	0.271	0.231	0.167	0.212
	RMSE	3.932	2.791	0.154	0.050	0.190
	RMSE%	0.307	0.356	0.295	0.219	0.272
	MBE	−0.018	−0.004	0.002	0.000	0.007
	MBE%	−0.001	−0.001	0.004	0.000	0.010
(e) S1Rand only	MAE	3.480	2.421	0.144	0.040	0.174
	MAE%	0.272	0.309	0.276	0.176	0.248
	RMSE	4.411	3.124	0.180	0.052	0.218
	RMSE%	0.344	0.399	0.346	0.229	0.312
	MBE	−0.066	−0.054	0.001	0.000	0.006
	MBE%	−0.005	−0.007	0.002	0.002	0.009

Table B.4
Comparison between our proposed method (upper table section, equivalent to the first section of Table 1) and a Random Forest baseline as it is frequently used for the estimation of vegetation parameters from remote sensing images (lower table section). The results indicate that in terms of MAE and RMSE (as well as their normalized counterparts), we outperform the baseline by a large margin for all structural variables.

		P95	MeanH	Dens	Gini	Cover
Ours	MAE	1.648	1.127	0.061	0.028	0.078
	MAE%	0.129	0.144	0.118	0.123	0.112
	RMSE	2.298	1.595	0.082	0.039	0.107
	RMSE%	0.179	0.204	0.158	0.170	0.154
	MBE	−0.086	−0.040	−0.003	−0.001	0.001
	MBE%	−0.007	−0.005	−0.006	−0.004	0.002
Random forest	MAE	3.124	2.195	0.123	0.039	0.148
	MAE%	0.244	0.280	0.237	0.173	0.212
	RMSE	4.051	2.893	0.158	0.052	0.192
	RMSE%	0.316	0.369	0.304	0.226	0.274
	MBE	0.111	0.095	−0.002	−0.001	0.000
	MBE%	0.009	0.012	−0.004	−0.005	0.000

References

Adnan, S., Maltamo, M., Coomes, D.A., García-Abril, A., Malhi, Y., Manzanera, J.A., Butt, N., Morecroft, M., Valbuena, R., 2019. A simple approach to forest structure classification using airborne laser scanning that can be adopted across bioregions. Forest Ecol. Manag. 433, 111–121.

Adnan, S., Maltamo, M., Mehtätalo, L., Ammaturo, R.N.L., Packalen, P., Valbuena, R., 2021. Determining maximum entropy in 3D remote sensing height distributions and using it to improve aboveground biomass modelling via stratification. Remote Sens. Environ. 260, 112464.

Alves de Almeida, D.R., Stark, S.C., Silva, C.A., Hamamura, C., Valbuena, R., 2020. URL: https://rdocumentation.org/packages/leafR/versions/0.3.

Alzate-Mejía, N., Santos-Boada, G., de Almeida-Amazonas, J.R., 2021. Decision-Making under uncertainty for the deployment of future hyperconnected networks: A survey. Sensors 21 (11).

Ashukha, A., Lyzhov, A., Molchanov, D., Vetrov, D., 2020. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In: International Conference on Learning Representations.

Astola, H., Seitsonen, L., Halme, E., Molinier, M., Lönnqvist, A., 2021. Deep neural networks with transfer learning for forest variable estimation using Sentinel-2 imagery in boreal forest. Remote Sens. 13 (12).

Astrup, R., Rahlf, J., Bjrkelo, K., Debella-Gilo, M., Gjertsen, A.-K., Breidenbach, J., 2019. Forest information at multiple scales: development, evaluation and application of the Norwegian forest resources map SR16. Scand. J. For. Res. 34 (6), 484–496.

Bergen, K., Goetz, S., Dubayah, R., Henebry, G., Hunsaker, C., Imhoff, M., Nelson, R., Parker, G., Radeloff, V., 2009. Remote sensing of vegetation 3-D structure for biodiversity and habitat: Review and implications for lidar and radar spaceborne missions. J. Geophys. Res.: Biogeosci. 114 (G2).

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural network. In: Proceedings of the 32nd International Conference on Machine Learning, Vol. 37. pp. 1613–1622.

Breidenbach, J., Waser, L.T., Debella-Gilo, M., Schumacher, J., Rahlf, J., Hauglin, M., Puliti, S., Astrup, R., 2021. National mapping and estimation of forest area by dominant tree species using Sentinel-2 data. Can. J. Forest Res. 51 (3), 365–379.

Calders, K., Adams, J., Armston, J., Bartholomeus, H., Bauwens, S., Bentley, L.P., Chave, J., Danson, F.M., Demol, M., Disney, M., et al., 2020. Terrestrial laser scanning in forest ecology: Expanding the horizon. Remote Sens. Environ. 251, 112102.

Carrasco, D., Diaz, J., Broquetas Ibars, A., Arbiol, R., Castillo, M., Pala, V., 1997. Ascending-descending orbit combination SAR interferometry assessment. In: Third ERS Symposium on Space At the Service of Our Environment: Florence, Italy: 14-21 March 1997. ESA Publications Division, pp. 1789–1794.

Caye Daudt, R., Le Saux, B., Boulch, A., 2018. Fully convolutional siamese networks for change detection. In: 25th IEEE International Conference on Image Processing (ICIP). pp. 4063–4067.

Chen, T., Fox, E., Guestrin, C., 2014. Stochastic gradient Hamiltonian Monte Carlo. In: Proceedings of the 31st International Conference on Machine Learning, Vol. 32. pp. 1683–1691.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2016. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. CoRR, abs/1606.00915.

Coops, N.C., Tompaski, P., Nijland, W., Rickbeil, G.J., Nielsen, S.E., Bater, C.W., Stadt, J.J., 2016. A forest structure habitat index based on airborne laser scanning data. Ecol. Indic. 67, 346–357.

Council, N.R., 2006. Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions using Weather and Climate Forecasts. The National Academies Press, Washington, DC.

de Lutio, R., D'Aronco, S., Wegner, J.D., Schindler, K., 2019. Guided super-resolution as pixel-to-pixel transformation. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8828–8836.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., et al., 2020. The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. Sci. Remote Sens. 1, 100002.

European Space Agency, 2020a. Sentinel Application Platform (SNAP) v8.0.0. [Online]. Available from: https://step.esa.int/main/download/snap-download/, Accessed: 2021-08-31.

European Space Agency, 2020b. SNAP Sentinel-1 Toolbox (S1TBX) v8.0. [Online]. Available from: https://sentinel.esa.int/web/sentinel/toolboxes/sentinel-1, Accessed: 2021-08-31.

European Space Agency, 2021a. ESA Copernicus Earth observation programme. [Online]. Available from: https://www.copernicus.eu/en, Accessed: 2021-08-31.

European Space Agency, 2021b. Sentinel-1 User Guide. [Online]. Available from: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-1-sar, Accessed: 2021-08-31.

FU Berlin, 2019. Remote Sensing Data Analysis online course. [Online]. Available from: https://blogs.fu-berlin.de/reseda/, Accessed: 2021-08-31.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning, Vol. 48. pp. 1050–1059.

Garnot, V.S.F., Landrieu, L., Giordano, S., Chehata, N., 2019. Time-space tradeoff in deep learning models for crop classification on satellite multi-spectral image time series. In: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 6247–6250.

Gast, J., Roth, S., 2018. Lightweight probabilistic deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3369–3378.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 580–587.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, http://www.deeplearningbook.org.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, Vol. 70. pp. 1321–1330.

Gustafsson, F.K., Danelljan, M., Schon, T.B., 2020. Evaluating scalable bayesian deep learning methods for robust computer vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1289–1298.

Hansen, M.C., Potapov, P.V., Goetz, S.J., Turubanova, S., Tyukavina, A., Krylov, A., Kommareddy, A., Egorov, A., 2016. Mapping tree height distributions in Sub-Saharan Africa using Landsat 7 and 8 data. Remote Sens. Environ. 185, 221–232.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Science 342 (6160), 850–853.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778.

Helber, P., Bischke, B., Dengel, A., Borth, D., 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens..

Hyyppä, J., Kelle, O., Lehikoinen, M., Inkinen, M., 2001. A segmentation-based method to retrieve stem volume estimates from 3-d tree height models produced by laser scanners. IEEE Trans. Geosci. Remote Sens. 39 (5), 969–975.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning, Vol. 37. PMLR, pp. 448–456.

Kaiser, P., Wegner, J., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning Aerial Image Segmentation from Online Maps. IEEE Trans. Geosci. Remote Sens. PP, 1–15.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 5580–5590.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations.

Knox, R.G., Peet, R.K., Christensen, N.L., 1989. Population dynamics in Loblolly Pine stands: Changes in skewness and size inequality. Ecology 70 (4), 1153–1167.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, Vol. 25.

Kuwata, K., Shibasaki, R., 2015. Estimating crop yields with deep learning and remotely sensed data. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). pp. 858–861.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, Vol. 30.

Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., Schindler, K., 2018. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. ISPRS J. Photogramm. Remote Sens. 146, 305–319.

Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., Wegner, J.D., 2022. Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. Remote Sens. Environ. 268, 112760.

Lang, N., Schindler, K., Wegner, J.D., 2019. Country-wide high-resolution vegetation height mapping with Sentinel-2. Remote Sens. Environ. 233, 111347.

Lang, N., Schindler, K., Wegner, J.D., 2021. High carbon stock mapping at large scale with optical satellite imagery and spaceborne LIDAR. arXiv preprint arXiv:2107.07431.

Laurin, G.V., Balling, J., Corona, P., Mattioli, W., Papale, D., Puletti, N., Rizzo, M., Truckenbrodt, J., Urban, M., 2018. Above-ground biomass prediction by Sentinel-1 multitemporal data in central Italy with integration of ALOS2 and Sentinel-2 data. J. Appl. Remote Sens. 12 (1), 1–18.

Lennart Noordermeer, T.G., Næsset, E., 2018. Direct and indirect site index determination for Norway Spruce and Scots Pine using bitemporal airborne laser scanner data. Forest Ecol. Manag. 428, 104–114.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440.

Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., Gascon, F., 2017. Sen2Cor for Sentinel-2. In: Bruzzone, L., Bovolo, F., Benediktsson, J.A. (Eds.), Image and Signal Processing for Remote Sensing XXIII.

Marmanis, D., Schindler, K., Wegner, J., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: improving semantic image segmentation with boundary detection. ISPRS J. Photogramm. Remote Sens. 135, 158–172.

Martin, D.M., Johnson, F.A., 2019. Incorporating uncertainty and risk into decision making to reduce nitrogen inputs to impaired waters. J. Environ. Manag. 249, 109380.

Næsset, E., 2002. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sens. Environ. 80 (1), 88–99.

Næsset, E., 2007. Airborne laser scanning as a method in operational forest inventory: Status of accuracy assessments accomplished in Scandinavia. Scand. J. For. Res. 22 (5), 433–442.

Næsset, E., Gobakken, T., Holmgren, J., Hyyppä, H., Hyyppä, J., Maltamo, M., Nilsson, M., Olsson, H., Persson, Å., Söderman, U., 2004. Laser scanning of forest resources: the nordic experience. Scand. J. For. Res. 19 (6), 482–499.

Neal, R.M., 1996. Bayesian Learning for Neural Networks. Springer-Verlag, Berlin, Heidelberg.

Nilsson, M., 1996. Estimation of tree heights and stand volume using an airborne lidar system. Remote Sens. Environ. 56 (1), 1–7.
Nilsson, M., Nordkvist, K., Jonzén, J., Lindgren, N., Axensten, P., Wallerman, J., Egberth, M., Larsson, S., Nilsson, L., Eriksson, J., Olsson, H., 2017. A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the National Forest Inventory. Remote Sens. Environ. 194, 447–454.
Nord-Larsen, T., Schumacher, J., 2012. Estimation of forest resources from a country wide laser scanning survey and national forest inventory data. Remote Sens. Environ. 119, 148–157.
Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems, Vol. 32.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch. In: NeurIPS Automatic Differentiation Workshop.
Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C.E., et al., 2021. Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sens. Environ. 253, 112165.
Potapov, P., Tyukavina, A., Turubanova, S., Talero, Y., Hernandez-Serna, A., Hansen, M., Saah, D., Tenneson, K., Poortinga, A., Aekakkararungroj, A., et al., 2019. Annual continuous fields of woody vegetation structure in the Lower Mekong region from 2000–2017 Landsat time-series. Remote Sens. Environ. 232, 111278.
Puliti, S., Breidenbach, J., Schumacher, J., Hauglin, M., Klingenberg, T.F., Astrup, R., 2021. Above-ground biomass change estimation using national forest inventory data with Sentinel-2 and Landsat. Remote Sens. Environ..
Puliti, S., Hauglin, M., Breidenbach, J., Montesano, P., Neigh, C., Rahlf, J., Solberg, S., Klingenberg, T.F., Astrup, R., 2020. Modelling above-ground biomass stock over Norway using national forest inventory data with ArcticDEM and Sentinel-2 data. Remote Sens. Environ. 236, 111501.
Ranganath, R., Gerrish, S., Blei, D., 2014. Black box variational inference. J. Mach. Learn. Res. 814–822.
Rasmussen, C.E., 2004. Gaussian processes in machine learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (Eds.), Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, TÜBingen, Germany, August 4 - 16, 2003, Revised Lectures. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 63–71.
Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788.
Rodríguez, A.C., D'Aronco, S., Schindler, K., Wegner, J.D., 2021. Mapping oil palm density at country scale: An active learning approach. Remote Sens. Environ. 261, 112479.
Rodríguez, A., Wegner, J., 2018. Counting the uncountable: deep semantic density estimation from Space. In: German Conference on Pattern Recognition. pp. 351–362.
Rußwurm, M., Körner, M., 2018. Multi-temporal land cover classification with sequential recurrent encoders. ISPRS Int. J. Geo-Inf. 7 (4), 129.
Rußwurm, M., Körner, M., 2020. Self-attention for raw optical satellite time series classification. ISPRS J. Photogramm. Remote Sens. 169, 421–435.
Rustowicz, R., Cheong, R., Wang, L., Ermon, S., Burke, M., Lobell, D., 2019. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 75–82.
Schneider, F.D., Morsdorf, F., Schmid, B., Petchey, O.L., Hueni, A., Schimel, D.S., Schaepman, M.E., 2017. Mapping functional diversity from remotely sensed morphological and physiological forest traits. Nature Commun. 8 (1), 1–12.
Senf, C., Mori, A., Müller, R., 2020. The response of canopy height diversity to natural disturbances in two temperate forest landscapes. Landsc. Ecol. 35, 2101–2112.
Senf, C., Pflugmacher, Y., Sebald, J., Knorn, J., Neumann, M., Hostert, P., Seidl, R., 2018. Canopy mortality has doubled in Europe's temperate forests over the last three decades. Nature Commun. 9, 4978.
Senf, C., Seidl, R., 2022. Post-disturbance canopy recovery and the resilience of Europe's forests. Global Ecol. Biogeogr. 31 (1), 25–36.
Shimizu, K., Ota, T., Mizoue, N., Saito, H., 2020. Comparison of multi-temporal PlanetScope data with Landsat 8 and Sentinel-2 data for estimating airborne LiDAR derived canopy height in temperate forests. Remote Sens. 12 (11).
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR.
Small, D., Holecz, F., Nuesch, D., 1995. Combination of ascending/descending ERS-1 InSAR data for calibration and validation. In: 1995 International Geoscience and Remote Sensing Symposium, IGARSS '95. Quantitative Remote Sensing for Science and Applications, Vol. 1. pp. 553–555, vol.1.
Sniazhko, S., 2019. Uncertainty in decision-making: A review of the international business literature. In: Muralidharan, E. (Ed.), Cogent Bus. Manage. 6 (1), 1650692.
Soroudi, A., Amraee, T., 2013. Decision making under uncertainty in energy systems: State of the art. Renew. Sustain. Energy Rev. 28, 376–384.
Spies, T.A., 1998. Forest structure: a key to the ecosystem. Northwest Sci. 72, 34–36.
Strutz, T., 2010. Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and beyond. Vieweg and Teubner, Wiesbaden, DEU.
Svein Solberg, H.K., Puliti, S., 2019. Age-independent site index mapping with repeated single-tree airborne laser scanning. Scand. J. For. Res. 34, 763–770.
Szegedy, C., Toshev, A., Erhan, D., 2013. Deep neural networks for object detection. In: Advances in Neural Information Processing Systems (NIPS), Vol. 26.
Turkoglu, M.O., D'Aronco, S., Perich, G., Liebisch, F., Streit, C., Schindler, K., Wegner, J.D., 2021a. Crop mapping from image time series: deep learning with multi-scale label hierarchies. arXiv preprint arXiv:2102.08820.
Turkoglu, M.O., D'Aronco, S., Wegner, J., Schindler, K., 2021b. Gating revisited: Deep multi-layer RNNs that can be trained. IEEE Trans. Pattern Anal. Mach. Intell..
Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E., Steininger, M., 2003. Remote sensing for biodiversity science and conservation. Trends Ecol. Evol. 18 (6), 306–314.
Tyukavina, A., Baccini, A., Hansen, M., Potapov, P., Stehman, S., Houghton, R., Krylov, A., Turubanova, S., Goetz, S., 2015. Aboveground carbon loss in natural and managed tropical forests from 2000 to 2012. Environ. Res. Lett. 10 (7), 074002.
Valbuena, R., Maltamo, M., Mehtätalo, L., Packalen, P., 2017. Key structural features of boreal forests may be detected directly using L-moments from airborne lidar data. Remote Sens. Environ. 194, 437–446.
Valbuena, R., O'Connor, B., Zellweger, F., Simonson, W., Vihervaara, P., Maltamo, M., Silva, C., Almeida, D., Danks, F., Morsdorf, F., Chirici, G., Lucas, R., Coomes, D., Coops, N., 2020. Standardizing ecosystem morphological traits from 3D information sources. Trends Ecol. Evol. 35 (8), 656–667.
Wang, L., Scott, K.A., Xu, L., Clausi, D.A., 2016. Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: A case study. IEEE Trans. Geosci. Remote Sens. 54 (8), 4524–4533.
Welling, M., Teh, Y.W., 2011. Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th International Conference on International Conference on Machine Learning. ICML '11, pp. 681–688.
Werbos, P.J., 1982. Applications of advances in nonlinear sensitivity analysis. In: Drenick, R.F., Kozin, F. (Eds.), System Modeling and Optimization.
Wilson, A.G., Izmailov, P., 2020. Bayesian deep learning and a probabilistic perspective of generalization. In: Advances in Neural Information Processing Systems, Vol. 33.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5987–5995.
Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2019. Joint deep learning for land cover and land use classification. Remote Sens. Environ. 221, 173–187.
Zheng, Z., Zeng, Y., Schneider, F.D., Zhao, Y., Zhao, D., Schmid, B., Schaepman, M.E., Morsdorf, F., 2021. Mapping functional diversity using individual tree-based morphological and physiological traits in a subtropical forest. Remote Sens. Environ. 252, 112170.