A Learnable Model With Calibrated Uncertainty Quantification for Estimating Canopy Height From Spaceborne Sequential Imagery

Leonidas Alagialoglou, *Student Member, IEEE*, Ioannis Manakos, Marco Heurich, Jaroslav Červenka, and Anastasios Delopoulos, *Member, IEEE*

***Abstract*— Global-scale canopy height mapping is an important tool for ecosystem monitoring and sustainable forest management. Various studies have demonstrated the ability to estimate canopy height from a single spaceborne multispectral image using end-to-end learning techniques. In addition to texture information of a single-shot image, our study exploits multitemporal information of image sequences to improve estimation accuracy. We adopt a convolutional variant of a long short-term memory (LSTM) model for canopy height estimation from multitemporal instances of Sentinel-2 products. Furthermore, we utilize the deep ensembles technique for meaningful uncertainty estimation on the predictions and postprocessing isotonic regression model for calibrating them. Our lightweight model (∼320k trainable parameters) achieves the mean absolute error (MAE) of 1.29 m in a European test area of 79 km^2. It outperforms the state-of-the-art methods based on single-shot spaceborne images as well as costly airborne images while providing additional confidence maps that are shown to be well calibrated. Moreover, the trained model is shown to be transferable in a different country of Europe using a fine-tuning area of as low as ∼2 km^2 with MAE = 1.94 m.**

***Index Terms*— Calibration, canopy height estimation, multitemporal regression, recurrent neural network (RNN), Sentinel-2, uncertainty estimation.**

I. Introduction

CHARACTERIZATION of 3-D forest structure (i.e., "the organization of the above-ground components of vegetation in space and time") is a fundamental step toward understanding the effects of climate change on ecosystem dynamics as well as sustainable forest management [1]. As a major component of forest structure, canopy height is a valuable indicator in numerous applications and research efforts, such as capturing above- and below-ground biomass patterns of ecosystems [2] or assessing and monitoring the health status of forest and its inhabitants [3], [4].

The standard method for canopy height model (CHM) estimation, apart from manual site inspection, is based on airborne LiDAR sensors that yield measured 3-D point clouds with ground sampling distance (GSD) <1 m and accuracy that, for our purposes, can be considered as ground truth. Due to the high cost of airborne measurements, they can only be applied at local scale without repetition. On the other hand, the two-year Global Ecosystem Dynamics Investigation (GEDI) mission has provided repetitive worldwide LiDAR measurements of the surface of the Earth from space [5]. However, the 25-m GSD of the spaceborne LiDAR sensor remains a limiting factor that needs to be addressed. Other techniques include aerial stereo imaging, such as the country-wide vegetation height mapping that was created by Ginzler and Hobi [6] or synthetic aperture radar (SAR) data [7].

Various approaches incorporate machine learning techniques in order to infer CHM from other sources of imagery. A supervised deep learning model for CHM estimation from single-shot Sentinel-2 imagery was introduced by Lang *et al.* [8]. A convolutional neural network (CNN) model, which is based on the Xception architecture [9], follows a preprocessing step for atmospheric correction and estimates canopy height maps at 10-m pixel resolution. The model has been trained and tested separately in two different datasets in Switzerland and Gabon. The root-mean-squared error (RMSE) of 3.4–5.6 m and the mean absolute error (MAE) between 1.7 and 4.3 m were achieved in pixel-wise evaluation. The authors argue that this low error in the tropics as well as in central Europe is suitable for country-scale canopy height mapping in terms of generalization and computation time.

Our prior work [10] evaluated the performance of a convolutional encoder–decoder network in the same forest area between Germany and Czech Republic. Apart from pixel-wise evaluation, aggregations of pixels based on their vegetation type were classified into six vegetation height classes. In the study of Boutsoukis *et al.* [11], SVM-based classification of extracted features of texture was used to classify land objects in vegetation classes of the general habitat category taxonomy

Manuscript received November 19, 2021; revised April 16, 2022; accepted April 25, 2022. Date of publication May 4, 2022; date of current version June 1, 2022. This work was supported by the European Union's Horizon 2020 Research and Innovation Program under Agreement 869520, NEXTLAND (https://ec-nextland.eu/). *(Corresponding author: Leonidas Alagialoglou.)*

Leonidas Alagialoglou and Anastasios Delopoulos are with the Multimedia Understanding Group, Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece (e-mail: lalagial@mug.ee.auth.gr).

Ioannis Manakos is with the Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), 57001 Thessaloniki, Greece.

Marco Heurich is with the Department of Visitor Management and National Park Monitoring, Bavarian Forest National Park, 94481 Grafenau, Germany, also with the Chair of Wildlife Ecology and Management, Faculty of Environment and Natural Resources, University of Freiburg, 79106 Freiburg, Germany, and also with the Institute for Forest and Wildlife Management, Inland Norway University of Applied Sciences, Campus Evenstad, 2480 Koppang, Norway.

Jaroslav Červenka is with the Šumava National Park, 34192 Kašperské Hory, Czech Republic.

Digital Object Identifier 10.1109/TGRS.2022.3171407

based on high-resolution airborne images on GSD smaller than 1 m.

An important research study that introduced the encoder–decoder deep architecture to regress height values from single-shot aerial images focused on mapping urban landscapes [12]. Similarly, the ISPRS benchmark datasets of Vaihingen and Potsdam have enabled other studies to explore end-to-end architectures that utilize single-shot remotely sensed imagery of urban environments [13].

Following the success of deep learning, remote sensing has widely adopted end-to-end learning solutions [14]. However, most approaches focus on the spatial and spectral dimension of the data, while only a few studies investigate the information content of temporal dimension contained in the abundance of available sequential imagery in the current Earth observation status. Change detection studies have developed an increasing number of algorithms with a well-established exploitation of multitemporal Earth observations [15]. In general, sequential data have been exploited with remarkable success in domains such as time series forecasting and natural language processing, by using recurrent neural network (RNN)-based approaches.

A clear demonstration of the superiority of multitemporal model over standard nontemporal models (CNNs and SVMs) for the land cover classification problem is presented in the work of Rußwurm and Körner [16]. Furthermore, in another work of the same authors [17], the effectiveness of multitemporal land cover classification with respect to cloud filtering is pointed out, which eliminates the need of any further preprocessing step. Our study is based on a similar approach to these studies, i.e., a convolutional variant of long short-term memory (LSTM) network, ConvLSTM, that was first introduced by Shi *et al.* [18] and [19]. A further adaptation of the model to a convolutional variant of gated recurrent (GRU) networks was presented in [17], although yielding comparable results with ConvLSTM architecture.

Another study toward incorporating sequential data investigates the use of multitemporal PlanetScope CubeSat to estimate canopy height in a pixel-wise manner using a random forest algorithm [20]. Furthermore, a combination of optical, Sentinel-2, with radar, Sentinel-1, and sequential data has been explored in the task of buildings' height estimation [21]. Several spectral–temporal and spatiotemporal features were calculated to train an SVM model that has been tested on a national scale.

An interesting interpretation of the valuable information contained in sequential spaceborne imagery has been given more than 25 years ago based on the first Landsat missions in the work of Odenweller and Johnson [22]. In this early work, individual crop types could be separated by manually inspecting the temporal profiles of a feature calculated from multiple bands indicating green vegetation. The reflectance levels in each crop type during seasons yield a distinctive shape of the temporal–spectral profiles. In addition, other studies demonstrate the predictive power of CHMs for land cover mapping [23].

In general, estimating height of land objects from multispectral image signal as captured from a satellite instrument is considered an ambiguous and ill-posed problem [12]. However, an indirect correlation is assumed of spectral signature with height that might be sufficient for straightforward applications. In specific, we assume that local spectral signatures are determined by a large vector of parameters that include shadowing, tree type, vegetation density, atmospheric column information, and so on, and some of these parameters are proxies to canopy height. Our intention is to exploit this indirect relation to the spectral signatures in order to predict canopy height. Moreover, it is important to make clear that our model is not trying to predict canopy height only from a single pixel's spectral signatures, rather to exploit the spatial distribution of these signatures as captured by CNNs, which inherently capture the corresponding texture of the satellite images.

In this article, we adopt an efficient spatiotemporal LSTM-based learning framework for the per-pixel CHM estimation, using sequences of Sentinel-2 products. With this model architecture, we are seeking to capture the temporal evolution of the distribution that the image texture follows and map it to the space of vegetation heights. Training and testing the model is based on different tiles of a total forest area of approximately 94 000 hectares, and comparison results with state-of-the-art studies [8], [10], [11] are provided.

Furthermore, an important aspect of our work is motivated by the need for quantifying the uncertainty of the model's predictions. In practical settings, knowing the confidence of the estimated CHM is useful in identifying the need for additional training data or ignoring low confident estimations. In this direction, various methods have been proposed, including deep ensembles [24], approximate Bayesian neural networks [25], and Monte Carlo dropout [26]. Since this is an open research area, we investigate the use of deep ensembles for meaningful confidence estimation, which is shown to outperform other methods in the literature [24].

Confidence in modern neural networks is shown to be poorly calibrated, i.e., probability estimates to be representative of the true correctness likelihood [27]. To calibrate our confidence maps, we train a second model that is agnostic to the main CHM model. Similar to the surprisingly successful method of "temperature scaling" [27], which is based on the entropy maximization principle for calibrating classification models, we implemented a simple calibration model based on the isotonic regression [28].

Finally, experimentation on transferability in time and location is performed. Reasonable height estimates are inferred for the years, 2018–2021, following the reference year, 2017, in the same geographic region. The model is also tested on a region in Switzerland, which is a different country that the training, allowing for a direct comparison with the state-of-the-art study of [8], as well as investigation of the size of fine-tuning dataset required. We concluded that a ground-truth area of ~ 2 km^2 of the Swiss region is sufficient to transfer our model with similar performance as in [8].

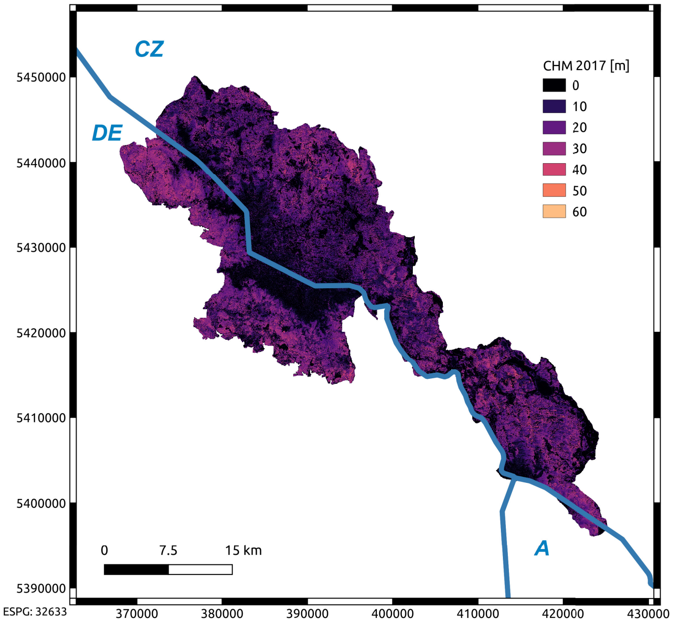

Fig. 1. Ground-truth canopy height model of the Bohemian Forest acquired from LiDAR measurement that took place in June 2017.

In short, the contributions of our study include the following models.

1) A model architecture for CHM estimation is suggested, based on sequences of satellite images without any cloud filtering. We evaluate its performance and compare with the state-of-the-art results based on single-shot approaches.
2) Alongside the CHM estimation model, we provide a method for meaningful and calibrated uncertainty quantification.
3) Experiments on transferability in location demonstrate that the model can be transferred in different regions with very limited fine-tuning dataset.
4) The impact of input sequence length in estimation accuracy is assessed.
5) Analysis of estimation error is performed based on terrain's characteristics, such as slope and aspect of the study area as well as cloud coverage.

II. Materials

A. Bohemian Forest (BF), Germany and Czech Republic

For training and testing our model a study area of 942 km^2 of the Bohemian Forest (BF) ecosystem was used. It is located at the borders between southeastern Germany and Czech Republic and includes two national parks, Bavarian Forest National Park and Šumava National Park. The forest area comprises heavily forested mountains with altitudes ranging between 570 and 1453 m. At higher altitudes, the landscape is covered with snow for 7–8 months, while in the valleys, snow persists for 5–6 months. Dominant tree species in the Bavarian area include Norway spruce (*Picea abies*), silver fir (*Abies alba*), and European beech (*Fagus sylvatica*) [29].

Ground-truth CHM of the study area was acquired from LiDAR measurements with the Riegl 680i sensor in June 2017 and is shown in Fig. 1. Details on the acquisition settings can be found in [11]. The GSD of the acquired 3-D point cloud is 1 m and the calculated CHM was bilinearly downsampled to 10-m resolution. For the preparation of the dataset, GDAL library and LAStools software with academic license were used.

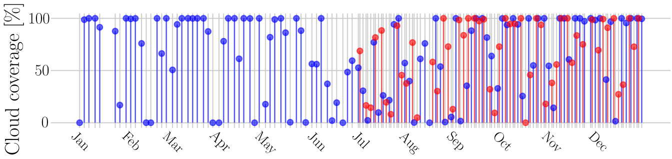

Fig. 2. Cloud coverage on all product dates of 2017. With blue color, S2A products are marked, and with red color, S2B products are marked. In total, 160 timeframes are available from the Sentinel-2 mission. The satellite S2B began providing data in the last semester of the year.

A sequence of Sentinel-2 Level-1C products, representing top-of-atmosphere reflectances, has been acquired for the study area during the year 2017, from the European Space Agency's (ESA) Copernicus Hub. As shown in Fig. 2, the complete sequence of products in 2017 consists of 160 timeframes in total from both satellites of the mission, S2A (blue) and S2B (red). Cloud coverage percentage for each product is also given in the same figure. In the years following 2017, more products are provided yearly for most areas since both satellites. In order to test for the transferability of the model through different years, the sequence of Sentinel-2 Level-1C has been acquired for a smaller part of the testing region, ~40 km^2, for the years 2018–2021. A random date sampling mechanism weighted in the second semester of each year was applied in the input sequences to simulate the products' frequency of year 2017. As discussed in the following section, no cloud coverage filter is applied in any of the Sentinel-2 sequences.

Finally, a land cover map of the area has been used for evaluation purposes and specifically in comparison with [10] and [11]. Based on the manual delineation of landscape patches (objects) [30], as created by local experts in 2012, we calculate aggregated pixel values of height in order to compare object-wise accuracy with the previous works.

B. Switzerland, CH

To investigate the model's transferability in different geographic locations, ground-truth vegetation height data are used for the region in Switzerland, CH, that is shown in Fig. 3. The vegetation height maps were calculated from stereo aerial imagery with a resolution of 1 × 1 m, based on photogrammetric image matching [6]. The transferability study area dataset was provided by the Swiss Federal Institute for Forest, Snow and Land-scape (WSL) for the year 2016. Similar to the BF ground-truth dataset, the vegetation height map is bilinearly downsampled to 10-m GSD. Following the reasoning of Lang *et al.* [8] that uses the same dataset, we filter height values >40 m as the only preprocessing step and the resulting map is considered accurate enough for the purposes of our work. The testing areas were selected with the minimum pixels of lakes to eliminate their influence on evaluation results.

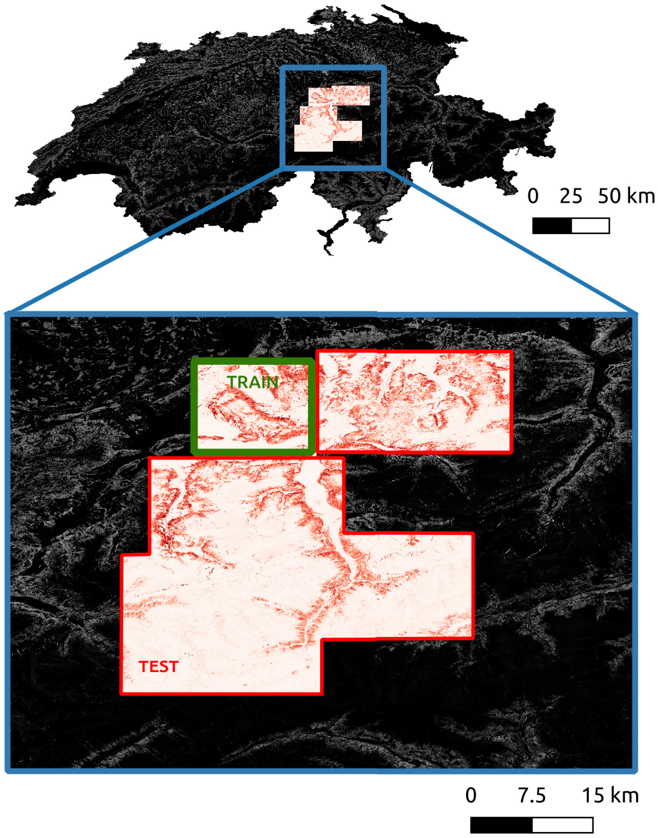

Fig. 3. Ground truth in Switzerland for year 2016 is used to evaluate the model's transferability in different geographic locations with limited or no training data. Regions in red are used for testing, while parts of the green region are used for fine-tuning model's parameters.

From the whole Switzerland area, a subregion of 2200 km^2 was selected, which corresponds to one of the two regions mentioned in [8], namely, CH2. In Fig. 3, regions in red color span ~2200 km^2 and are used for testing purposes, while parts of the green region, which cover a total area of ~320 km^2, are used for fine-tuning model's parameters. In fact, it is argued that only a small portion of the whole testing area is needed for fine-tuning.

The sequences of Sentinel-2 Level-1C products for the described areas have been downloaded for the whole year 2016. In total, 67 Sentinel-2 products were available during that year.

III. Methodology

A. Model Architecture

In our regression problem, we infer a parametric model $f : \mathbf{x} \in \mathbb{R}^{B\times T\times w\times h} \mapsto \{\hat{\mathbf{y}} \in \mathbb{R}^{w\times h}, \hat{\mathbf{s}} \in \mathbb{R}^{w\times h}\}$. The input $\mathbf{x}$ is a multitemporal, multispectral tile of size $w \times h$ pixels with B bands for each pixel in T distinct dates. The model maps the input to an estimation map, $\hat{\mathbf{y}}$ of the real height map $\mathbf{y}$, as well as an estimation of the prediction error, $\hat{\mathbf{s}}$, also called confidence map.

We adopt a convolutional variant of LSTM network, as introduced by Xingjian *et al.* [18] for the precipitation forecasting problem and further investigated by Patraucean *et al.* [19] in a weakly supervised framework for semantic segmentation of videos. Building on the important advantages of LSTM networks over simpler RNN approaches (i.e., dealing with the vanishing gradient problem as well as capturing long-term temporal dependencies), the ConvLSTM cell enables capturing of spatiotemporal correlations for our per-pixel regression problem.

The complete model, named thereafter `spatioTempCHM`, utilizes a number of ConvLSTM cells and is shown in Fig. 4. The input tile is a sequence of Sentinel-2 L1C images $\mathbf{x} = \{x_t\}_{t=1}^{T}$ and consists of $T = 40$ timeframes that have been selected randomly from the available products of the year 2017, as shown in Fig. 2. Each timeframe, x_t, with size 48 × 48 pixels includes 13 bands and is fed into a 2-D convolutional layer with kernel size 3 × 3 and output depth of 64, resulting in an input feature map, $\mathcal{X}_t$ at time t. A ConvLSTM cell, as described next, is applied for each feature map together with the cell state, $\mathcal{C}_t$ and hidden state, $\mathcal{H}_t$ of the previous timeframe. The first timeframe is initialized with tensors $\mathcal{C}_0 = \mathbf{0}, \mathcal{H}_0 = \mathbf{0}$ as the previous cell and hidden states. The cell state of the last timeframe is convolved with kernel size 3 × 3 and output depth of 32. Finally, two separate fully connected layers for each pixel are applied for the computation of the output values that represent the estimated height values and estimated prediction errors. All tensors prior to convolutional layers are zero-padded with a single pixel in the edges to maintain the input size.

The ConvLSTM cell replaces the standard full connections of input-to-state and state-to-state transitions of LSTM with the convolution operator. The key equations of the ConvLSTM cell architecture as adapted by Patraucean *et al.* [19] are given next

$$\begin{aligned}
i_t &= \sigma(w_{xi} * \mathcal{X}_t + w_{hi} * \mathcal{H}_{t-1} + b_i) \quad \text{(input gate)}\\
f_t &= \sigma\left(w_{xf} * \mathcal{X}_t + w_{hf} * \mathcal{H}_{t-1} + b_f\right) \quad \text{(forget gate)}\\
o_t &= \sigma(w_{xo} * \mathcal{X}_t + w_{ho} * \mathcal{H}_{t-1} + b_o) \quad \text{(output gate)}\\
\mathcal{H}_t &= o_t \circ \tanh(\mathcal{C}_t) \quad \text{(hidden state)}\\
\mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ \tanh(w_{xc} * \mathcal{X}_t + w_{hc} * \mathcal{H}_{t-1} + b_c)\\
&\quad \text{(cell state)}
\end{aligned}$$

where $\mathcal{X}_t$ is the input feature map at time t; $\mathcal{C}_t$ and $\mathcal{H}_t$ are the outputs of the ConvLSTM cell, named cell state and hidden state, respectively; and w_{xi}, w_{xf}, w_{xo}, w_{xc}, w_{hi}, w_{hf}, w_{ho}, and w_{hc} and b_i, b_f, b_o, and b_c represent the trainable 3 × 3 convolution kernels and biases.[1] The intermediate variables, i_t, f_t, and o_t, represent the input, forget, and output gates, respectively. The sigmoid σ and tanh are used as activation functions, introducing nonlinearities. Convolution operation and element-wise multiplication are represented by $*$ and $\circ$, respectively.

B. Uncertainty Quantification

The standard cost function for training the model with a single output, i.e., only $\hat{\mathbf{y}}$, is mean squared error (MSE)

[1] The input convolutional kernel that maps the input x_t to $\mathcal{X}_t$ could be merged with the kernels w_{xi}, w_{xf}, w_{xo}, and w_{xc} because they are consecutive convolutions without nonlinearities in between. However, separating convolutional layers reduces total trainable parameters with no cost at model's performance.

between the estimated mean and target height values. In this case, the cost function is optimized with respect to the model parameters, θ, in batches of 16 tiles, with Adam optimizer and learning rate 10^{-3}.

However, in order to further acquire a confidence map, that expresses the data uncertainty of each prediction, we assume a Gaussian target error distribution and train the model based on the Gaussian negative log-likelihood (GNLL) cost function [31]. Based on this assumption, we represent the model's outputs with the estimated mean $\hat{\boldsymbol{y}}$ and the estimated variance $\hat{\boldsymbol{\sigma}}^2$ of canopy height. In fact, for numerical reasons [25], the network is trained to output the log variance of a pixel i, $\hat{s}_i := \log \hat{\sigma_i}^2$. In this case, the cost function is optimized using rms-Prop optimizer in batches of 16 tiles with learning rate and weight decay 10^{-4}. For D the number of output pixels, the GNLL cost function takes the form

$$\mathcal{L}_{\text{GNLL}}(\theta) = \frac{1}{D}\sum_{i}^{D}\frac{1}{2}\left(e^{-\hat{s}_i}||y_i - \hat{y}_i||^2 + \hat{s}_i\right). \tag{1}$$

Apart from the heteroscedastic aleatoric uncertainty, which captures noise inherent in the collected data, we wish to capture the epistemic uncertainty as well, which describes our ignorance about the model. For this purpose, deep ensembles technique [24] is utilized, which has been shown to outperform other methods, such as approximate Bayesian NNs [24] or Monte Carlo dropout [26]. We obtain N ensemble models for the prediction of mean and variance $\{\hat{\mathbf{y}}_n, \hat{\boldsymbol{\sigma}_n}^2\}_{n=1}^N$, trained on the same training and validation set but with different initial parameters. The outputs of the ensemble models for each pixel $(\hat{y}_n, \hat{\sigma}_n)$ are then combined as a mixture of Gaussians, thus

$$\hat{y} = \frac{1}{N}\sum_{n}^{N}\hat{y}_n \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n}^{N}\left(\hat{\sigma}_n^2 + \hat{y}_n^2\right) - \hat{y}^2. \tag{3}$$

We can rearrange the formula as

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}\hat{\sigma}_n^2 + \frac{1}{N}\sum_{n=1}^{N}\left(\hat{y}_n^2 - \hat{y}^2\right) \tag{4}$$

which is equivalent as adding the mean value of the variances σ_n^2 with the variance of the means μ_n of the Gaussian distribution members. In the literature [25], the first term expresses the aleatoric uncertainty, while the second term expresses the epistemic uncertainty.

C. Confidence Calibration

Even with the deep ensembles technique, we observed that the predictions were not well calibrated; in fact, they were underconfident. As a final step for calibrating the output confidence map, a second model that is agnostic to the main `spatioTempCHM` model was trained on the confidence predictions of the validation set based on the isotonic regression model [28]. This small and simple model is used for post-processing the estimated confidence (variance) of the test set. Despite its simplicity, this method, similar to the temperature scaling in classification tasks [27], generalizes the uncertainty predictions on the test set surprisingly well [32].

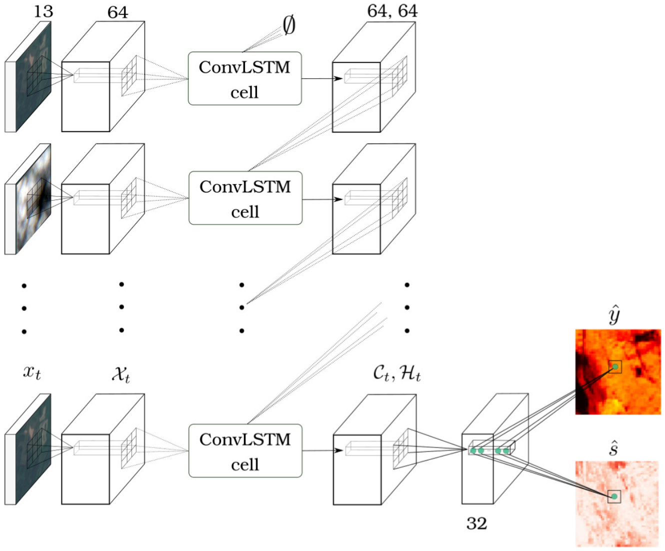

Fig. 4. Unfolded recurrent neural network architecture with ConvLSTM cells for canopy height estimation with uncertainty quantification. A sequence of 40 Sentinel-2 L1C tiles is used as input, $\{x_t\}_{t=1}^T$, while the output consists of two maps with the same size as input representing estimated height mean value, $\hat{y}$ and log variance, $\hat{s}$.

The most common metric for quantifying calibrations is the expected calibration error (ECE). If we group the confidence predictions into M bins of equal intervals and B_m is the set of indices of pixels whose confidence prediction falls into the mth interval, then

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{L}|\text{err}(B_m) - \text{std}(B_m)| \tag{5}$$

where $\text{err}(B_m) = (1/|B_m|)\sum_{i\in B_m}(|y_i - \hat{y}_i|)$, L number of pixels, and $\text{std}(B_m) = (1/|B_m|)\sum_{i\in B_m}\hat{\sigma}_i$.

D. Experimental Details

The BF study area is divided into nonoverlapping square tiles with 48 pixels on each side. Tiles at the borders of the study area that included at least one pixel with no-data value have been discarded. The 80% (7.03 Mpixels corresponding to 703 km^2) of the total number of tiles is used for training, while 10% (0.786 Mpixels corresponding to 78.6 km^2) is used as validation set during training for hyperparameter selection and stop training epoch. The rest 10% of the tiles (0.790 Mpixels corresponding to 79 km^2) are used for testing purposes and performance evaluation. This main separation in the BF dataset is based on the random selection of distinct subregions for train/validation/test and is called random split. Another separation of the dataset that is mentioned later in this section is based on the geographic location by manually selecting the southeast 72 km^2 of the whole area for testing and the rest for train/validation. It is named location-based-split.

A total number of $N = 6$ `spatioTempCHM` ensemble models have been trained and evaluated in terms of accuracy and confidence calibration. Input data to the model consist of

40 products randomly selected during the year 2017, as shown in Fig. 2, without any cloud coverage filtering. The isotonic regression model is trained on the 10% validation set and is used as a postprocessing step for calibrating the confidence map.

Each `spatioTempCHM` model of the ensemble has ~320 k trainable parameters, which is considered a very small neural network compared to other common architectures for the same kernel size, such as `SegNet` with ~30 M parameters, `U-Net` ~31 M parameters, or `ConvEnc-Dec` from our previous work [10] with ~19 M parameters. A NVIDIA RTX 2080Ti card was used for training with a total training time of approximately 15 h. Inference, although based on ensembling, is relatively quick and can be easily deployed in large-scale web applications. In specific, fetching the input data of ~300 km^2 and 40 timeframes to memory is the most time-consuming task, with ~1.5-min duration, while the actual inference in GPU lasts ~20 s for a single ensemble member.

Estimation accuracy is evaluated pixel-wise by calculating RMSE and MAE of the mean height value for all test pixels that cover a total area of 79 km^2. The state-of-the-art results of Lang *et al.* [8] and our previous work [10] are compared with the resulted RMSE and MAE.

To compare accuracy with the study of Boutsoukis *et al.* [11] and our previous work [10], we used an object-wise testing method that is described in detail in [11] and [10]. Both studies involve the same study area as ours. Land objects are considered as aggregation of pixels based on the land cover map of the area. For each delineated object, an average height value was calculated from all pixels in it. In order to adapt comparison with the classification scheme mentioned in [11], the estimated height value of each land object is quantized in four classes of height [11], representing a subset of general habitat categories. The percentage of correctly classified land objects (object-based accuracy) and the percentage of correctly classified area considering the area of each land object (area-based accuracy) are used as metrics for the comparison.

To provide a wider understanding on the strengths and weaknesses of `spatioTempCHM` model, we investigate the correlation of the terrain's characteristics with the estimation error. In specific, we analyzed correlations of the actual estimation error and the estimated confidence with slope and aspect for each pixel based on the LiDAR measurements. In a similar manner, the influence of clouds on the model's performance is investigated by conducting an analysis of estimation error with respect to the average cloud coverage of pixels in the sequence of 40 timeframes randomly selected during the reference year 2017. The estimated confidence is also included in the analysis, allowing for conclusions on the model's cloud robustness.

Another important aspect that was investigated for our multitemporal model is the impact of sequence length in the prediction accuracy. For this purpose, different ensemble models have been trained on smaller number of input images around the central date of June 1, 2017 and comparison results are shown in Fig. 9. The central date was selected based on the acquisition date of the LiDAR dataset (June 2017) as well as the available cloud-free products around this date, based on Fig. 2, in order to include enough cloud-free input data for the models with small input sequence length, e.g., 5. In this comparison, since we care about comparing solely the accuracy of each ensemble model, only the first output of the model architecture was used in training, representing the height value and omitting the confidence estimation; the cost function of MSE was used in this case.

An additional experiment was conducted to compare `spatioTempCHM` model with a simpler temporal aggregation strategy that does not exploit temporal features. The baseline model uses the encoder–decoder network, `ConvEnc-Dec`, that is described in our previous work [10], as a feature extractor. The 40 feature vectors for each pixel are averaged before feeding the aggregated vector in a fully connected layer to regress height values. This model, named `ConvEnc-Dec-mean40`, incorporates 40 timeframes around the central date but does not exploit their ordering. The model's parameters are initialized using the pretrained feature extractor on single-shot cloud-free images in our previous work. End-to-end retraining on the sequence of images contributes to the cloud robustness of the baseline method.

Estimated height and confidence maps for the years 2018–2021 of a subregion of 40-km^2 BF are calculated to indirectly evaluate transferability of the `spatioTempCHM` model in time.

Finally, we include further experimentation to address the question of model's transferability in geographic location. First, comparison results between location-based split and random split, as described in Section II, are informative on the transferability in regions of close proximity to the training set. Second, to properly evaluate transferability to region of a different country, we tested the trained model to a subregion of 2200 km^2 in Switzerland, as shown in Fig. 3. A simple fine-tuning process, without any frozen layer, is performed on regions of variable size to identify the smallest area that is sufficient to fine-tune the model with similar accuracy to the state of the art [8]. This comparison is possible since the same subregion of Switzerland is used in [8] and the same preprocessing steps are followed. The fine-tuning datasets consist of a number of 48 × 48 pixel tiles that are randomly selected from the fine-tuning area, shown with green in Fig. 3.

IV. Experimental Results

A. Accuracy and Calibration Results

The depicted randomly selected test tiles in Fig. 5 are of size 48 × 48 pixels with 10-m resolution. Assuming Gaussian target error distribution, the predicted mean, $\hat{\mathbf{y}}$, and standard deviation (std), $\hat{\boldsymbol{\sigma}}$, are inferred from the six-ensemble `spatioTempCHM` model, as described in Section III. The ground truth, $\mathbf{y}$, that is compared against the predicted mean value is measured using the airborne LiDAR sensory. The last column of Fig. 5 depicts the pixel-wise absolute error, $|\mathbf{y} - \hat{\mathbf{y}}|$. In Fig. 6, the predicted mean values of all test pixels are scattered against the target ground-truth values.

Quantitative evaluation is performed for the model's accuracy and calibration, i.e., how well the confidence map

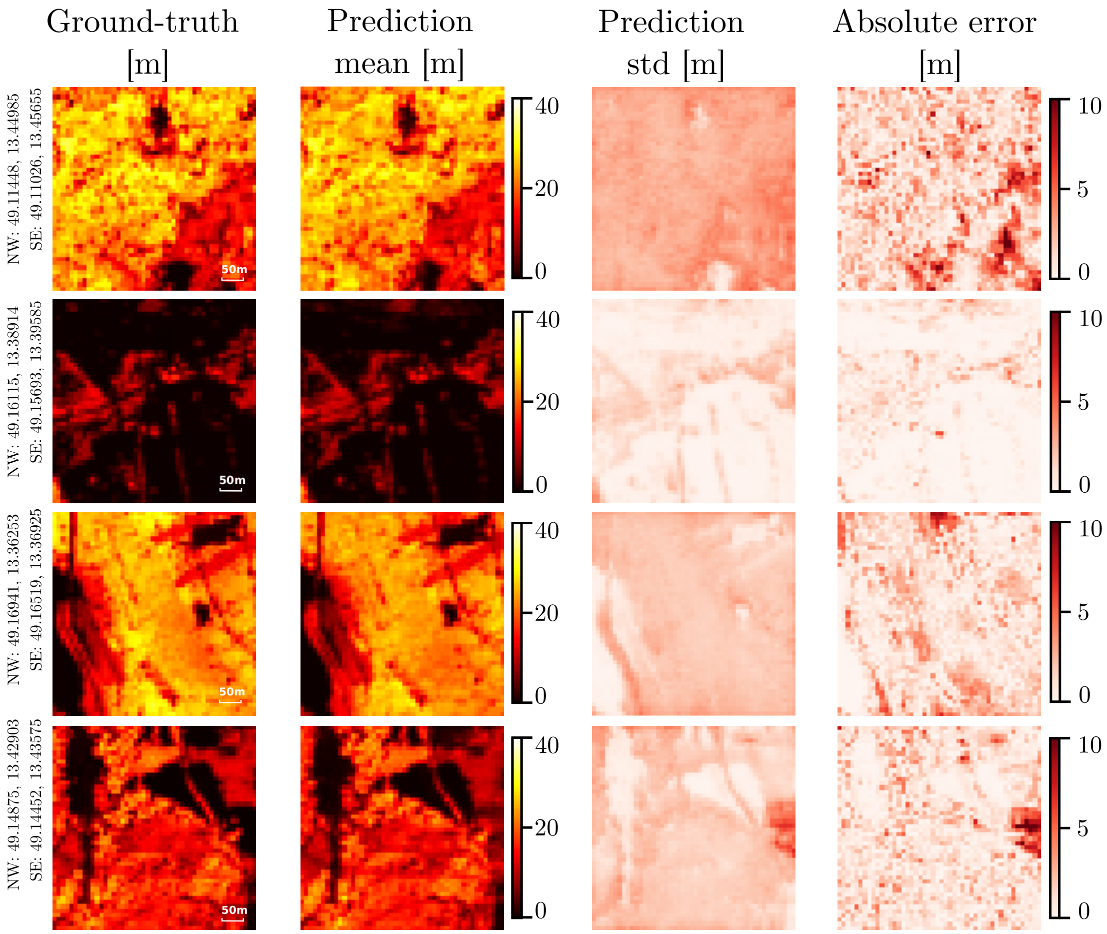

Fig. 5. Predicted mean and standard deviation of canopy height alongside with LiDAR measured ground truth and absolute error. A six-member ensemble of `spatioTempCHM` model is used for estimating the depicted tiles of 48 × 48 pixels with 10-m resolution.

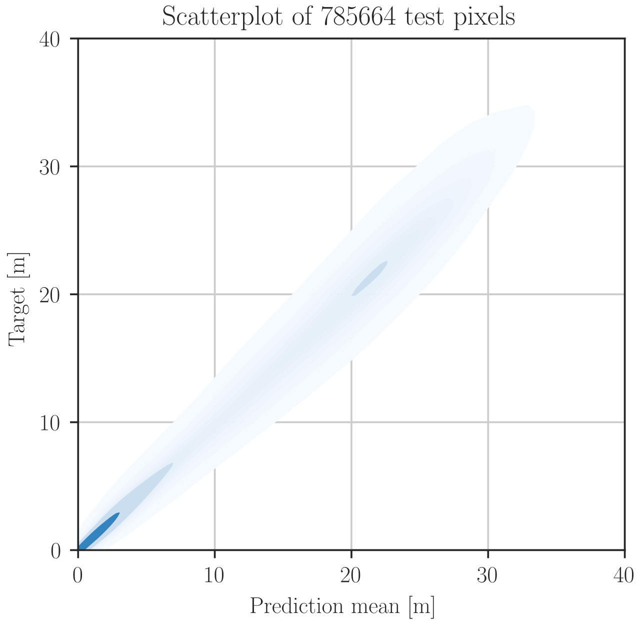

Fig. 6. Ground-truth versus estimated height values for more than 78 km^2 of test area.

TABLE I

PIXEL-WISE COMPARISON RESULTS OF `spatioTempCHM` MODEL WITH STATE OF THE ART

Method	Location	Area	MAE $[m]$	RMSE $[m]$
Lang et al. [8]	Switzerland	91Mpx	1.7	3.4
Lang et al. [8]	Gabon	25Mpx	4.3	5.6
`ConvEnc-Dec` [10]	BF	9.4Mpx	2.29	3.15
`ConvEnc-Dec-mean40`	BF	9.4Mpx	2.04	3.05
`spatioTempCHM`	BF	9.4Mpx	**1.29**	**1.87**

estimates the absolute error map. Resulted RMSE and MAE in pixel-wise evaluation are presented in Table I and next to the state-of-the-art results. While our method utilizes a sequence of Sentinel-2 products as input, the compared methods are considered as single-shot methods. However, the study of Lang *et al.* [8] selects 4–12 products filtered for cloud coverage of less than 70%. Similarly, our previous work [10] averages the pixel-wise estimation of three different products filtered for cloud coverage of less than 4%.

An object-wise testing method, as described in Section III, was used to compare the results with [11] and [10]. The same dataset area with the previous works of ~21 km^2 is used. In [11], different accuracy results are given for three types of testing objects based on their size (large, medium, and small), whereas in our previous [10] and current work, there is only one type of objects, regardless of their size. The object-based accuracy, i.e., percentage of correctly classified objects, and the area-based accuracy, i.e., percentage of correctly classified area, are given in Table II.

TABLE II

OBJECT-WISE COMPARISON RESULTS OF SIX-ENSEMBLE spatioTempCHM WITH [10] AND [11] IN FOUR-CLASS QUANTIZATION

Method	Number of class objects	Object-based acc(%)	Area-based acc(%)
Boutsoukis et al. [11]	Large: 90 Medium.: 1671 Small: 2006	91.11 80.73 66.55	91.39
ConvEnc-Dec [10]	2604	91.40	94.10
spatioTempCHM	2604	**95.74**	**98.21**

A common way for visual inspection of model calibration is the reliability diagram, as shown in Fig. 7, before (uncalibrated) and after (calibrated) the model-agnostic postprocessing model trained on the validation set and tested on the test set. In these diagrams, the accuracy, i.e., absolute error, is plotted against the estimated confidence, i.e., standard deviation. For clarity reasons, all test pixels are grouped into bins of the same size, with the diameter of each dot representing the power of the bin. If the model is perfectly calibrated, then the diagram should be the identity function (red dashed line). Regions of confidence below the perfect diagonal represent miscalibrated underconfident predictions, while regions above represent overconfident predictions. In both calibrated and uncalibrated models, ECE has been calculated for comparison (ECE $= 0$ corresponds to perfect calibration).

B. Analysis of Estimation Error

With respect to the characteristics of the terrain, the error analysis of the estimation is performed for the six-ensemble spatioTempCHM model. The slope of each location and its aspect are plotted in histogram bins against MAE in Fig. 8. The same error analysis is given in Fig. 8(c) with respect to the average cloud coverage of the input sequence.

C. Impact of Input Sequence Length

As a final step in our analysis, we assessed the impact of input sequence length on prediction accuracy, as described in Section III. Comparison results of the different ensemble models trained on a different number of input images around the central date of June 1, 2017 are shown in Fig. 9. The best performing ensemble model with five members (five-ensemble) with 40 timeframes' input, as shown in Fig. 9, resulted in MAE $=$ 1.33 m, whereas MAE $=$ 1.29 m is given in Table I. The reason for this minor discrepancy is the use of 40 timeframes around the central date in the former configuration, while in the latter configuration, 40 timeframes are selected randomly during the period of a year. The results of baseline model ConvEnc-Dec-mean40 that does not utilize temporal information are given in Table I and are also plotted in Fig. 9 (orange).

D. Transferability in Time

The Sentinel-2 time series of a region in BF with area $\sim$40 km^2 for the years, 2018–2021, following the reference year, 2017, as described in Section II, are used for estimating the canopy height map. The median values of canopy height for each year, together with their estimated error bars, are given in Fig. 10. The ground-truth value of the year 2017 is also shown with the red dot. These estimations are based on all datatakes throughout each year, but the estimated height is considered to represent the height in June because of the training set characteristics.

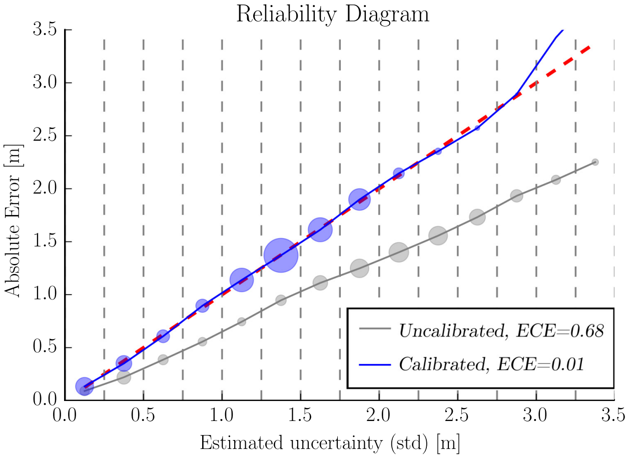

Fig. 7. Reliability diagram using six-ensemble spatioTempCHM model, before (uncalibrated) and after (calibrated) isotonic regression calibration technique. Diameter of dots represents the power (i.e., number of pixels) of each bin.

E. Transferability in Geographic Location

Performance evaluation of spatioTempCHM model in different regions is given in Table III. The spatioTempCHM model is trained in BF and fine-tuned using subregions in Switzerland of different sizes. To investigate on the impact of the selected "CH smallest" subregion, the fine-tuning was performed multiple times to allow for quantification of the variation in performance introduced due to this factor.

Although the spatioTempCHM architecture was trained on tiles of 48 × 48 pixels, there is no restriction on the tile dimensions since the network's mask can be applied in an infinitely large grid. Artifacts introduced by merging estimated tiles can be eliminated by increasing the tile size. For demonstration, the canopy height map of a large area in Switzerland is given in Fig. 11, next to high-resolution RGB satellite and aerial image, captured within 3–5 years [33].

In Fig. 12, we present the descriptive statistics of the uncertainty types in the estimated confidence maps of different geographic locations. Quantified uncertainty, as described in (4), is the sum of model (epistemic) and data (aleatoric) uncertainty. Histograms of the ratio between epistemic and aleatoric uncertainty are given in Fig. 12 for the four scenarios in rows 1–3 and 9 of Table III and represent different degrees of distribution shift due to geographic location. The percentage of pixels with ratio >1 is given in each scenario.

V. DISCUSSION

Visual inspection of the predicted mean in the testing tiles of Fig. 5 demonstrates the accurate CHM estimation

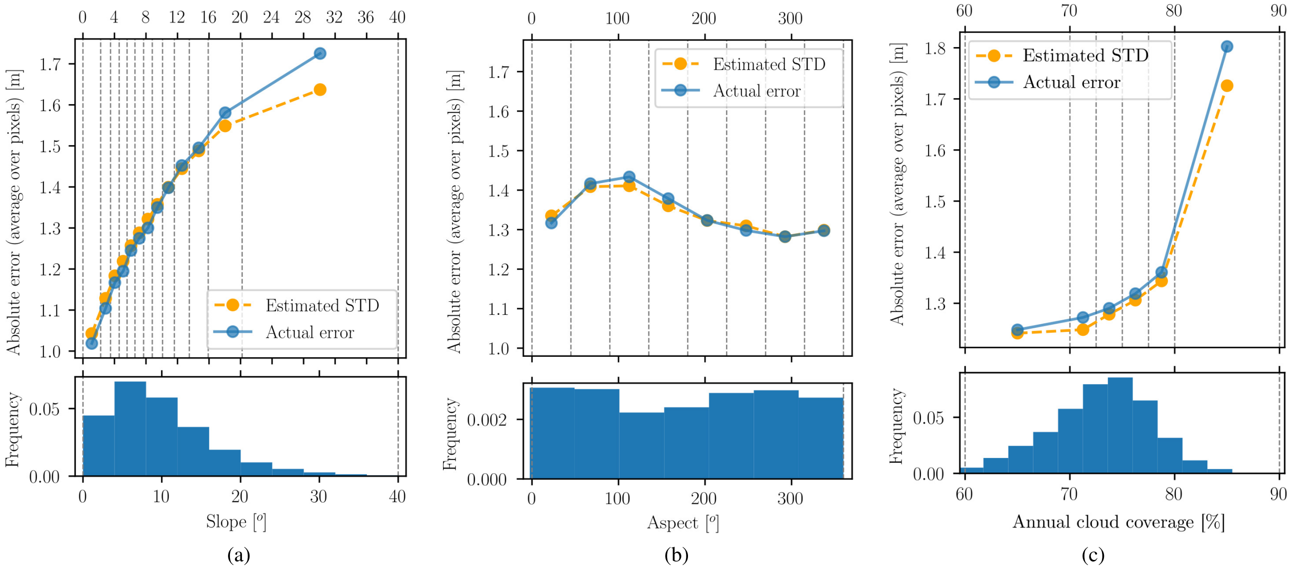

Fig. 8. Correlation of estimation error with slope, aspect, and average cloud coverage of pixels in the test area. The estimated uncertainty for each bin is given. (a) Error versus slope. (b) Error versus aspect. (c) Error versus annual cloud coverage.

TABLE III

PERFORMANCE EVALUATION OF spatioTempCHM MODEL IN DIFFERENT GEOGRAPHIC LOCATIONS

Location **Train → Test**	**Fine-tune area**	**Test area**	**MAE [m]**	**RMSE [m]**	**ECE [m]**	**ECE uncalibrated [m]**
BF → BF (random split)	no fine-tune	$79km^2$	1.29	1.87	0.01	0.68
BF → BF (location-based split)	no fine-tune	$72km^2$	1.76	2.50	0.09	0.44
BF + CH all $\xrightarrow{fine-tune}$CH	$320km^2$	$2200km^2$	1.52	2.92	0.30	0.47
BF + CH $1/8 \xrightarrow{fine-tune}$ CH	$40km^2$	$2200km^2$	1.49	2.99	0.38	0.51
BF + CH $1/16 \xrightarrow{fine-tune}$ CH	$20km^2$	$2200km^2$	1.57	3.07	0.31	0.66
BF + CH $1/32 \xrightarrow{fine-tune}$ CH	$10km^2$	$2200km^2$	1.65	3.24	0.30	0.29
BF + CH $1/64 \xrightarrow{fine-tune}$ CH	$5km^2$	$2200km^2$	1.69	3.26	0.31	0.40
BF + CH smallest $\xrightarrow{fine-tune}$ CH	$\mathbf{2.30km^2}$	$\mathbf{2200km^2}$	$\mathbf{1.94}(\mathbf{SD:.04})$	$\mathbf{3.83}(\mathbf{SD:.01})$	$\mathbf{0.56}(\mathbf{SD:.19})$	$\mathbf{0.52}(\mathbf{SD:.02})$
BF → CH	no fine-tune	$2200km^2$	2.60	4.16	- [2]	1.16

[2] Interestingly, if we calibrate the model by using as little as $0.23km^2$ of the Swiss region, the ECE drops significantly to $0.44m$.

in different forest areas with high and low vegetation. The meaningful and calibrated standard deviation estimation maps that accompany the estimated height maps provide a useful indication of our confidence in the predictions. For instance, in the middle right side of the last tile in Fig. 5, we observe a region of higher absolute error that is successfully captured by our confidence map.

Quantitatively, our model yielded an MAE of 1.29 m and an RMSE of 1.87 m in a test area of 79 km^2 in the same geographic location and date with the training dataset. Compared with the state-of-the-art results that are based on single-shot Sentinel-2 images, our model outperforms the highest MAE and RMSE of Lang *et al.* [8]. The work of [8] uses 4–12 timeframes filtered for low cloud coverage, but the temporal information of the sequential imagery is not exploited. Furthermore, object-wise evaluation according to the methodology of Boutsoukis *et al.* [11] yielded higher object- and area-based accuracies. However, the SVM approach mentioned in [11] might offer different advantages, for example, in terms of computation time.

In Fig. 9, comparison results of accuracy between models with different input sequence lengths indicate the positive correlation of the number of input images with accuracy. Longer input sequence lengths around a central date yield higher accuracy.

The aforementioned results comply with our intuition that the use of multitemporal observations improves accuracy in estimating the Earth's parameters, i.e., canopy height in our study. Similar conclusions that highlight the use of multitemporal instead of single-shot imagery are derived from the work of Rußwurm and Körner [17] in the problem of land cover classification. An important research question is whether the improved predictive power is due to the actual ordering of the sequential information or rather the larger quantity and variability of the bag-of-images is sufficient. Inquiring into this, we perform comparison with the baseline model ConvEnc-Dec-mean40, as shown in Fig. 9 and Table I. The performance drop from MAE from 1.29 to 2.04 m from our point of view is significant, supporting the claim that temporal information improves prediction accuracy. We find the idea of investigating the use of attention-based mechanism, with or without considering the ordering, interesting as a future work.

Regarding the confidence maps that accompany the estimated height maps, we accept that a meaningful uncertainty quantification, in terms of aleatoric and epistemic uncertainty, is provided by incorporating deep ensembles technique.

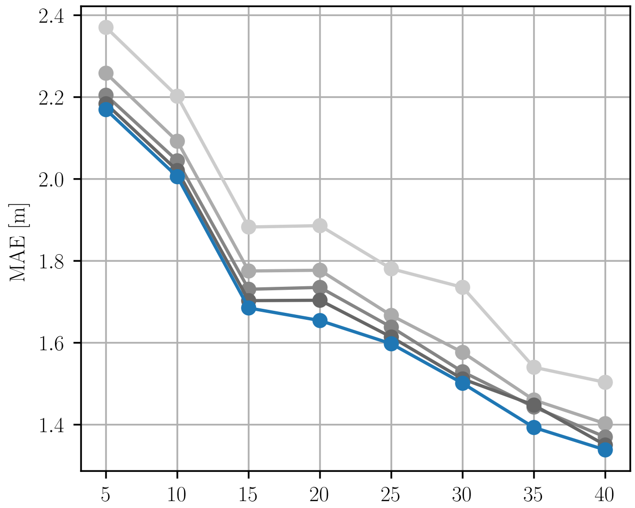

Fig. 9. Mean absolute error (MAE) of `spatioTempCHM` model for different input sequence lengths (T). Results with fewer ensemble members are shown with gray colors and results with reference point (∗) `ConvEnc-Dec-mean40` that does not utilize temporal information are shown in orange.

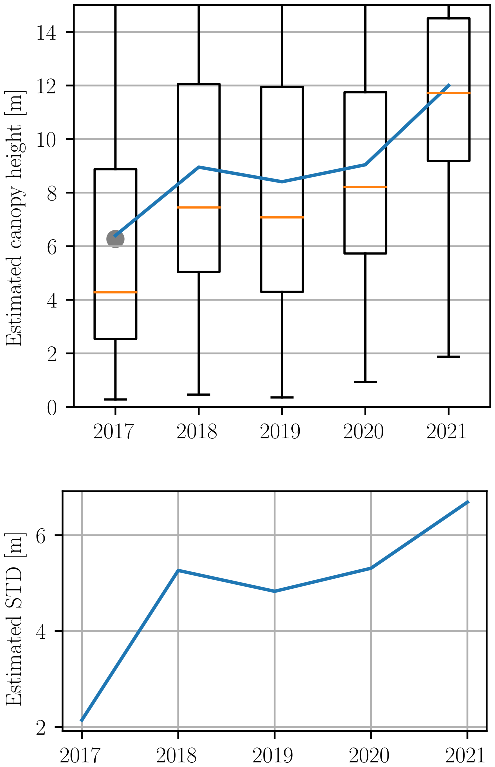

Fig. 10. (Top) Mean values (blue) and boxplots of canopy height for years following reference year, 2017, in a BF subregion of ~40 km^2, as calculated by the trained `spatioTempCHM` ensemble models. Ground-truth value of the year 2017 is shown with the gray dot. (Bottom) Quantified uncertainty, indicating the confidence in the height predictions, as estimated by `spatioTempCHM` for the corresponding years. Mean values over all pixels are given.

Calibration using the model-agnostic isotonic regression technique has been shown to reduce significantly the ECE, as shown in the reliability diagram of Fig. 7. Interesting insights on the distinction between the estimated data (aleatoric) and model (epistemic) uncertainty are presented in Fig. 12. The experimental findings indicate that transition to new domains with different input distributions infects significantly the ratio between the two types of estimated uncertainty.

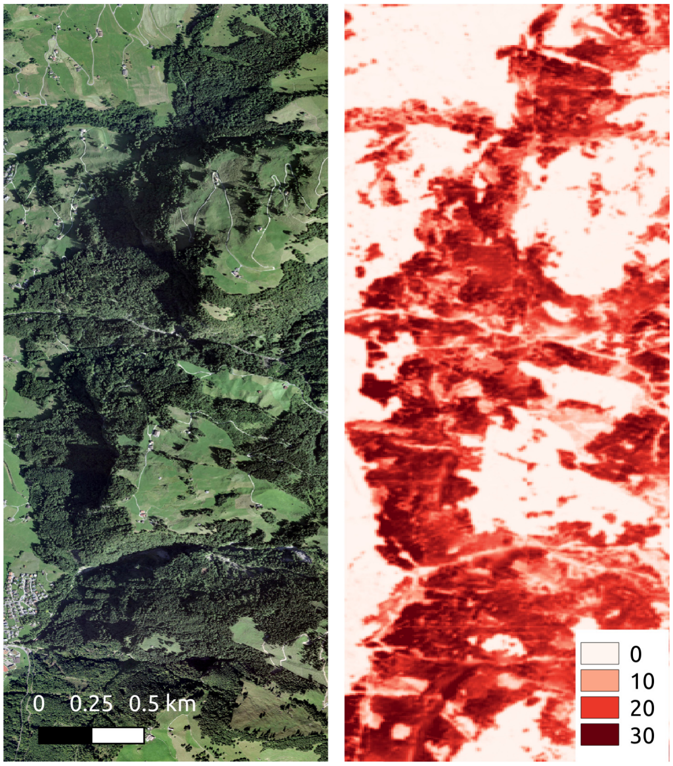

Fig. 11. Estimated canopy height map of a large area in Switzerland next to high-resolution RGB image [33]. Colorbar units in meters.

Based on the error analysis, we observe in Fig. 8(a) and (b) that the estimation error appears significantly correlated with the slope of the terrain, while the aspect of each location explains minor variance of the estimation error and needs further investigation.

Furthermore, we found a positive correlation of estimation error with the average cloud coverage of the input sequence, as shown in Fig. 8(c). This demonstrates the cloud robustness of the method and combined with the correlation of the confidence with cloud coverage suggests that we have a tool for identifying and eventually discarding uncertain predictions due to high average cloud coverage in the sequence of satellite images. For interpreting the way cloudy observations are handled by the network, we refer to the experiments of Rußwurm and Körner [34] for cloud robustness of ConvLSTM networks, indicating that the model learns a cloud-filtering mechanism without any training for this specific task.

In Section I, we discussed the reasons for assuming that the Sentinel-2 time series can fit the canopy height maps of a region and this can be useful for downstream applications. However, the question of whether this can be achieved reliably is equivalent to which factors affect the generalizability of the model. In an effort to address this question, we evaluated the transferability of the trained model in time and geographic location.

In Fig. 10, the estimated canopy height maps for the years following the reference year, 2017, are compared. We observe that the estimated height maps are within a reasonable range. Another criterion for evaluating performance in the absence of

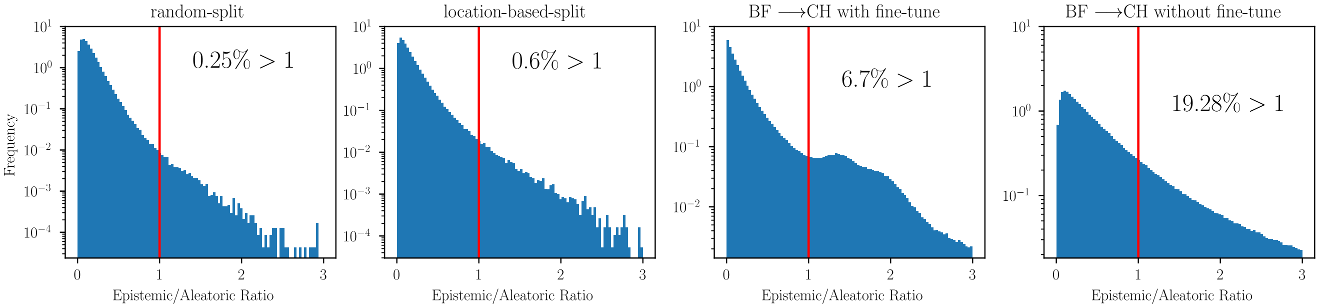

Fig. 12. Quantified uncertainty, as described in 4, consists of model (epistemic) and data (aleatoric) uncertainty. Histograms of the ratio between epistemic and aleatoric uncertainty, estimated in different geographic locations (out-of-distribution). The four histograms correspond to rows 1–3 and 9 of Table III and represent different degrees of distribution shift due to geographic location. The percentage of pixels with ratio >1 is given in each histogram.

ground truth is based on the realistic expectation of gradual canopy height increase year after year. The results seem to follow this expected trend in general. However, there seems to be a higher estimated uncertainty in the year 2018 and an abnormal decrease in height during the season 2018–2019. A mere speculative explanation can be given by the fact that, to the extent of our knowledge, 2018 has been a significantly drier year for that region.

Model transferability in different regions is examined using a subregion of 2200 km^2 in Switzerland of 2016, which corresponds to the region of [8], characterized as CH2. Apart from model transfer investigation, a direct comparison with state-of-the-art results is possible since the same preprocessing steps are applied. In Table III, we present the performance evaluation of the model in different geographic locations than the training dataset without or with very limited fine-tuning dataset. Surprisingly, we conclude that a ground-truth area of ~2 km^2 of the Swiss region is sufficient to transfer our model with similar performance accuracy as the state-of-the-art study of Lang *et al.* [8] that uses a corresponding training area of ~2700 km^2. Our model achieved MAE = 1.94 m after fine-tuning with ~2 km^2, which is comparable with MAE = 2 m in [8].

To investigate the impact of the specific "CH smallest" subregion that is selected, fine-tuning was performed multiple times with randomly selected subregions to allow for quantification of the variation in performance introduced due to this factor. It appears that it has limited influence in prediction accuracy, but the calibrated ECE demonstrates a slightly higher standard deviation.

Exploring the transferability of such models in different geographic locations and time, as well as other possible factors, e.g., tree type and growth state, is a crucial step toward robust application deployment. Apart from simply fine-tuning the model, as we applied, more sophisticated techniques, such as domain adaptation methods or few-shot learning [35], are worth investigating. We believe that our model is a good basis for such techniques, due to its relatively small parameter dimensionality.

VI. Conclusion

This study proposes a neural network architecture that provides accurate canopy height maps from multitemporal spaceborne imagery alongside with a method for estimating meaningful and calibrated confidence maps. The resulted MAE of 1.29 m, based on 40 timeframes' Sentinel-2 images in a test area of 79 km^2, outperforms the state-of-the-art results of single-shot input approaches.

The results suggest that higher estimation accuracy can be achieved by incorporating the widely available sequences of satellite images compared to single-shot approaches. Furthermore, based on our relatively lightweight network architecture, ~320k trainable parameters, we conclude that the significantly improved estimation accuracy does not come at the cost of computation time.

Quantifying the confidence of the estimated CHM can be a useful tool in practical settings by identifying the need for additional training data or by ignoring low confident estimations. We adopt the use of deep ensembles technique for meaningful uncertainty quantification, while a postprocessing isotonic regression model yielded calibrated confidence maps.

An estimation error analysis showed that the estimation error is larger in steeper slopes, but minor correlation was observed with the aspect of the pixel. Similarly, the positive correlation of cloud coverage with confidence estimates demonstrates the method's robustness in cloud coverage. Furthermore, by investigating the effect of input sequence length, we conclude that the longer the sequence length around a central date, the higher the accuracy is achieved.

Finally, experiments on transferability in time and geographic location reveal the potential uses of the lightweight model in practical settings. Reasonable height estimates that demonstrate generally increasing trend are inferred for the years, 2018–2021, following the reference year, 2017. Transferring the model in a different country of Europe proved to perform surprisingly well, especially after fine-tuning with as little as a ~2 km^2 of ground-truth area, which yielded similar performance with state-of-the-art model, trained on ~2700-km^2 ground-truth area.

Acknowledgment

The authors would like to acknowledge the support of the "Data Pool Initiative for the Bohemian Forest Ecosystem" data-sharing initiative of the Bavarian Forest National Park. LiDAR data were granted by the Cross-border cooperation program Czech Republic–Bavaria Free State ETC goal 2014–2020, the Interreg V Project No. 99 "Přeshranični mapovani lesnich ekosystemů—cesta ke společnemu managementu NP Šumava a NP Bavorsky les/Grenzuberschreitende Kartierung der Waldokosysteme—Weg zum gemeinsamen Management in NP Sumava und NP Bayerischen Wald." Moreover, they also acknowledge the support of Christian Ginzler and the Swiss Federal Institute for Forest, Snow and Landscape (WSL) for providing vegetation height data for Switzerland in the framework of the Swiss National Forest Inventory (NFI). Finally, they also thank the European Space Agency for sponsoring their EDC Sentinel Hub subscription.

References

[1] M. Pardini *et al.*, "Early lessons on combining lidar and multi-baseline SAR measurements for forest structure characterization," *Surv. Geophys.*, vol. 40, no. 4, pp. 803–837, Jul. 2019.

[2] X. Wang, S. Ouyang, O. J. Sun, and J. Fang, "Forest biomass patterns across northeast China are strongly shaped by forest height," *Forest Ecol. Manage.*, vol. 293, pp. 149–160, Apr. 2013, doi: 10.1016/j.foreco.2013.01.001.

[3] S. Goetz, D. Steinberg, R. Dubayah, and B. Blair, "Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA," *Remote Sens. Environ.*, vol. 108, no. 3, pp. 254–263, Jun. 2007.

[4] A. L. Mitchell, A. Rosenqvist, and B. Mora, "Current remote sensing approaches to monitoring forest degradation in support of countries measurement, reporting and verification (MRV) systems for REDD+," *Carbon Balance Manage.*, vol. 12, no. 1, pp. 1–22, Dec. 2017.

[5] R. Dubayah *et al.*, "The global ecosystem dynamics investigation: High-resolution laser ranging of the Earth's forests and topography," *Sci. Remote Sens.*, vol. 1, Jun. 2020, Art. no. 100002.

[6] C. Ginzler and M. Hobi, "Countrywide stereo-image matching for updating digital surface models in the framework of the Swiss national forest inventory," *Remote Sens.*, vol. 7, no. 4, pp. 4343–4370, Apr. 2015.

[7] M. Recla and M. Schmitt, "Deep-learning-based single-image height reconstruction from very-high-resolution SAR intensity data," *J. Photogramm. Remote Sens.*, vol. 183, pp. 496–509, Jan. 2022.

[8] N. Lang, K. Schindler, and J. D. Wegner, "Country-wide high-resolution vegetation height mapping with Sentinel-2," 2019, *arXiv:1904.13270*.

[9] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise separable convolutions for neural machine translation," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–10.

[10] L. Alagialoglou, I. Manakos, M. Heurich, J. Červenka, and A. Delopoulos, *Canopy Height Estimation From Spaceborne Imagery Using Convolutional Encoder-Decoder* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12573. Berlin, Germany: Springer, 2021, pp. 307–317.

[11] C. Boutsoukis, I. Manakos, M. Heurich, and A. Delopoulos, "Canopy height estimation from single multispectral 2D airborne imagery using texture analysis and machine learning in structurally rich temperate forests," *Remote Sens.*, vol. 11, no. 23, p. 2853, Dec. 2019. [Online]. Available: http://www.mdpi.com/journal/remotesensing

[12] L. Mou and X. Xiang Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," 2018, *arXiv:1802.10249*.

[13] H. A. Amirkolaee and H. Arefi, "Height estimation from single aerial images using a deep convolutional encoder-decoder network," *J. Photogramm. Remote Sens.*, vol. 149, pp. 50–66, Mar. 2019.

[14] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[15] M. Molinier *et al.*, "Optical satellite image time series analysis for environment applications: From classical methods to deep learning and beyond," in *Change Detection and Image Time Series Analysis 2: Supervised Methods*. 2021, pp. 109–154.

[16] M. RuBwurm and M. Korner, "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 11–19.

[17] M. Rußwurm and M. Körner, "Multi-temporal land cover classification with sequential recurrent encoders," *Int. J. Geo-Inf.*, vol. 7, no. 4, p. 129, 2018.

[18] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[19] V. Patraucean, A. Handa, and R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory," 2015, *arXiv:1511.06309*.

[20] K. Shimizu, T. Ota, N. Mizoue, and H. Saito, "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.*, vol. 12, no. 11, p. 1876, Jun. 2020.

[21] D. Frantz *et al.*, "National-scale mapping of building height using Sentinel-1 and Sentinel-2 time series," *Remote Sens. Environ.*, vol. 252, Jan. 2021, Art. no. 112128.

[22] J. B. Odenweller and K. I. Johnson, "Crop identification using landsat temporal-spectral profiles," *Remote Sens. Environ.*, vol. 14, nos. 1–3, pp. 39–54, Jan. 1984.

[23] C. A. Mücher *et al.*, "Synergy of airborne LiDAR and worldview-2 satellite imagery for land cover and habitat mapping: A bio_sos-eodham case study for The Netherlands," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 37, pp. 48–55, May 2015.

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," 2016, *arXiv:1612.01474*.

[25] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp.1–11.

[26] J. Caldeira and B. Nord, "Deeply uncertain: Comparing methods of uncertainty quantification in deep learning algorithms," *Mach. Learn., Sci. Technol.*, vol. 2, no. 1, Jul. 2020, Art. no. 015002.

[27] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2017, pp. 1321–1330.

[28] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 625–632.

[29] M. Cailleret, M. Heurich, and H. Bugmann, "Reduction in browsing intensity may not compensate climate change effects on tree species composition in the bavarian forest national park," *Forest Ecol. Manage.*, vol. 328, pp. 179–192, Sep. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037811271400320X

[30] R. S. Gonzalez, H. Latifi, H. Weinacker, M. Dees, B. Koch, and M. Heurich, "Integrating LiDAR and high-resolution imagery for object-based mapping of forest habitats in a heterogeneous temperate forest landscape," *Int. J. Remote Sens.*, vol. 39, no. 23, pp. 8859–8884, Dec. 2018.

[31] D. A. Nix and A. S. Weigend, "Estimating the mean and variance of the target probability distribution," in *Proc. IEEE Int. Conf. Neural Netw.*, vol. 1. Jul. 1994, pp. 55–60. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/374138/

[32] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, vol. 6, 2018, pp. 4369–4377.

[33] *World Imagery [Basemap]*, ESRI, Redlands, CA, USA, 2022.

[34] M. Rußwurm and M. Körner, "Convolutional LSTMs for cloud-robust segmentation of remote sensing imagery," 2018, *arXiv:1811.02471*.

[35] M. Ruswurm, S. Wang, M. Korner, and D. Lobell, "Meta-learning for few-shot land cover classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 200–201.

Leonidas Alagialoglou (Student Member, IEEE) received the Diploma degree from the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece, in 2013, and the M.Sc. degree in sustainable agriculture from the School of Agriculture, AUTH, in 2020, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering.

He worked as an Embedded Software Engineer at Kenotom P.C., Kalamaria, Greece, for the development of large plant models used in hardware-in-the-loop simulators of the automotive industry. He is also a member of the Multimedia Understanding Group, AUTH, where he is working as a Research Associate in the EU-funded projects. His research interests include uncertainty quantification for deep learning, Bayesian modeling, remote sensing imaging, and statistics for spatiotemporal data.

Ioannis Manakos received the Ph.D. degree from the Technical University of Munich, Munich, Germany.

He has been a Principal Researcher in remote sensing with the Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece, since 2012. He was the Head of the "Geoinformation in Environmental Management M.Sc./Department" at the International Centre for Advanced Mediterranean Agronomic Studies for seven years. He has coordinated or participated in more than 45 European and national research, innovation, and development projects under various funding frameworks (including FP6, FP7, and H2020).

Dr. Manakos chaired the European Association of Remote Sensing Laboratories (EARSeL) (2012–2014) and is a Copernicus Academy Member.

Marco Heurich received the Ph.D. degree in forest remote sensing from the Technical University of Munich, Munich, Germany, in 2008.

He is currently the Head of the Department of National Park Monitoring, Bavarian Forest National Park, Grafenau, Germany, and a Professor of wildlife ecology and conservation biology with the University of Freiburg, Freiburg im Breisgau, Germany, and the Inland Norway University of Applied Sciences, Hamar, Norway. He has authored more than 221 ISI journal articles and 22 book chapters. His research interests include remote sensing and its applications in forest ecology and wildlife ecology and conservation.

Prof. Heurich received a scholarship for highly gifted students from Friedrich-Naumann-Stiftung and was a recipient of the Lennart-Bernadotte-Award for landscape ecology.

Jaroslav Červenka received the Ph.D. degree in forest ecology from the Czech University of Life Sciences in Prague, Prague, Czechia, in 2016.

He is currently the Head of the Forest Monitoring Department in Administration of the Šumava National Park. He is interested in the research of forest ecology, remote sensing, wildlife ecology, and nature conservation.

Anastasios Delopoulos (Member, IEEE) was born in Athens, Greece, in 1964. He received the bachelor's degree from the Department of Electrical Engineering, National Technical University of Athens (NTUA), Athens, Greece, in 1987, the M.Sc. degree from the University of Virginia, Charlottesville, VA, USA, in 1990, and the Ph.D. degree from NTUA, in 1993.

From 1995 to 2001, he was a Senior Researcher with the Institute of Communication and Computer Systems, NTUA. Since 2001, he has been with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece, where he is currently a Professor. He is the (co)author of more than 75 journal and conference scientific papers. His research interests lie in the areas of semantic analysis of multimedia data and computer vision. He has participated in 21 European and national research and development projects related to the application of signal, image, video, and information processing to entertainment, culture, education, and health sectors.

Dr. Delopoulos is a member of the Technical Chamber of Greece.