Remote Sensing of Environment 311 (2024) 114270

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Generation of country-scale canopy height maps over Gabon using deep learning and TanDEM-X InSAR data

Daniel Carcereri [a,b,*], Paola Rizzoli [a], Luca Dell'Amore [a], José-Luis Bueso-Bello [a], Dino Ienco [c,d], Lorenzo Bruzzone [b]

[a] German Aerospace Center (DLR), Microwaves and Radar Institute, Weßling, Germany
[b] Università degli Studi di Trento, Trento, Italy
[c] INRAE, TETIS, Université de Montpellier, Montpellier, France
[d] INRIA, Université de Montpellier, Montpellier, France

ARTICLE INFO

Edited by Jing M. Chen

Keywords:
Forest height
Forest parameter regression
Deep learning
Bistatic SAR
Interferometric coherence
InSAR
TanDEM-X
LVIS

ABSTRACT

Operational canopy height mapping at high resolution remains a challenging task at country-level. Most of the existing state-of-the-art inversion methods propose physically-based schemes which are specifically tuned for local scales. Only few approaches in the literature have attempted to produce country or global scale estimates, mostly by means of data-driven approaches and multi-spectral data sources. In this paper, we propose a robust deep learning approach that exploits single-pass interferometric TanDEM-X data to generate accurate forest height estimates from a single interferometric bistatic acquisition. The model development is driven by considerations on both the final performance and the trustworthiness of the model for large-scale deployment in the context of tropical forests. We train and test our model over the five tropical sites of the AfriSAR 2016 campaign, situated in the West Central state of Gabon, performing spatial cross-validation experiments to test its generalization capability. We define a specific training dataset and input predictors to develop a robust model for country-scale inference, by finding an optimal trade-off between the model performance and the large-scale reliability. The proposed model achieves an overall estimation bias of 0.12 m, a mean absolute error of 3.90 m, a root mean squared error of 5.08 m and a coefficient of determination of 0.77. Finally, we generate a time-tagged country-scale canopy height map of Gabon at 25 m resolution, discussing the potential and challenges of these kinds of products for their application in different scenarios and for the monitoring of forest changes.

1. Introduction

The regular and precise monitoring of the state of Earth's forests is of paramount importance for preservation efforts (FAO, 2020). The assessment of a forest's health, dynamics and resources can be achieved through the measurement and monitoring of proxy indices, such as the canopy height, the above-ground biomass density or the canopy cover fraction. Conventionally, the most precise way to estimate such forest variables is to acquire them manually on-ground on a per-tree basis, which is both time consuming and expensive (Picard et al., 2012; Jucker et al., 2017). In order to characterize forests on regional or national scales statistical acquisition strategies that approximate the area of interest with representative sampling grids are typically introduced (Bundeswaldinventur, 2024).

To achieve wall-to-wall estimates at large-scales, it is necessary to derive these parameters from satellite imagery, by relating the biophysical forest parameters to the acquired spaceborne feature-maps. This can be achieved by either inverting physical-based models, which attempt to describe the interaction of the transmitted signals with the forest structure, or by relying on data-driven approaches, which directly learn the underlying relationship from large-amounts of informative case samples.

In this scenario, Synthetic Aperture Radar (SAR) sensors have received great attention from the remote sensing (RS) community, as the interaction between the electromagnetic waves and the imaged scatterers strongly depends on the geometrical and the dielectric properties of the target, i.e., on the characteristics of the vegetation. Here, the Random Volume over Ground (RVoG) model probably represents one of the most studied and understood physical interpretations of the InSAR microwave interaction with the forest structure, characterizing the scattering profile of vegetation as the combination of a Dirac-like ground component and a vertical distribution of randomly

* Corresponding author at: German Aerospace Center (DLR), Microwaves and Radar Institute, Weßling, Germany.
E-mail address: daniel.carcereri@dlr.de (D. Carcereri).

https://doi.org/10.1016/j.rse.2024.114270
Received 23 January 2024; Received in revised form 29 May 2024; Accepted 12 June 2024
Available online 29 June 2024

oriented scatterers (Papathanassiou and Cloude, 2001; Cloude and Papathanassiou, 2003). The model inversion requires a single-baseline fully-polarimetric (i.e., quad-pol.) acquisition, allowing for the estimation of the forest height. Modern spacerbone SAR systems, such as the German TanDEM-X mission (Krieger et al., 2007, 2013), are indeed also capable of acquiring fully-polarimetric InSAR products, but these typically do not represent the operational acquisition mode for large-scale surveys, as they can only be acquired as experimental products over limited test sites. Therefore, much effort has gone into the definition of effective strategies to invert such a model in presence of non-fully polarimetric data (Chen et al., 2016; Olesk et al., 2016).

More recently, research attention has shifted towards sparse fusion strategies, aiming at exploiting the availability of modern dedicated spaceborne LiDAR-based missions (e.g., GEDI, ICESat-2) to retrieve the model parameters necessary for proper model inversion. In Denbina et al. (2018) it was proposed to train a support vector machine (SVM) with sparse LiDAR samples, in order to select the optimal baseline for the RVoG model inversion with NASA's UAVSAR (L-band) and LVIS instruments. The validation over the study areas of the AfriSAR Campaign (Fatoyinbo et al., 2021) achieved an RMSE varying between 4.99 m and 5.99 m. In Guliaev et al. (2021), the authors proposed to avoid the parametrization of the simplified vertical profile functions, and instead to estimate these directly from LiDAR waveforms. Forest height inversion from TanDEM-X coherence samples led to a root mean square error (RMSE) of 8.16 m and to a squared Pearson correlation coefficient r^2 of 0.16 over the AfriSAR test site of Lopé, after removing 10.69% of the estimates, which fell below the interferometric phase center. Finally, by generating two separate profiles for vegetation below 25 m and above 40 m, respectively, and interpolating the profiles in the transitional range, the authors achieved a RMSE of 8.62 m and a r^2 of 0.40 (after dropping 14.17% of underestimated samples). Similarly, in Choi et al. (2023) a mean TanDEM-X vertical reflectivity profile was estimated using the zero-order eigenvector of the diagonalized profile covariance matrix, derived from GEDI LiDAR waveforms. Using this approach, they generated a continuous 25 m forest height map of the island of Tasmania, Australia. In comparison with reference LiDAR measurements, their proposed approach achieved RMSE values between 6.6 m and 7.2 m, and r^2 values between 0.40 and 0.42, respectively, depending on the considered orbit direction and orthogonal baseline.

In the last two decades, data-driven approaches have seen a major surge for remote sensing applications, offering a completely different paradigm: instead of building up semi-empirical, physically-based models and retro-fitting them to the existing data, they take advantage of the availability of large quantities of heterogeneous data and learn the underlying relationships with physical phenomena from the data itself. Up to now, most of the works published in the literature mainly relied on the use of multi-spectral optical data.

In Potapov et al. (2021), the authors proposed a bagged regression tree ensemble-based approach to predict canopy heights at a resolution of 30 m from multi-temporal Landsat acquisitions on a global scale. The machine learning algorithm was calibrated per-image using GEDI-derived RH95 estimates as reference, to estimate forest height from a mixture of Landsat-derived features expressing spectral, phenological, statistical and temporal properties of the scene. They obtained a mean absolute error (MAE) of 6.36 m, an RMSE of 9.07 m and a coefficient of determination R^2 of 0.61.

More recently, Deep Learning (DL) approaches have received most of the attention, as these take advantage of local and non-local spatial patterns to improve the performance accuracy over less sophisticated pixel-wise approaches. In Lang et al. (2019) a fully convolutional approach, based on the Xception DL architecture, was proposed, which was trained to regress forest canopy height from multiple Sentinel-2 multi-spectral acquisitions. The authors trained and validated their approach both over two alpine regions in Switzerland using stereophotogrammetry-derived measurements, and over the five AfriSAR sites in Gabon using LVIS-derived measurements. Over Gabon, the authors obtained a MAE of 4.9 m and an RMSE of 6.5 m, respectively, using the least clouded acquisitions, and of 4.3 m and 5.6 m when considering the median height. In Becker et al. (2023), the authors proposed a Bayesian Deep Learning (BDL) approach which was validated at country-level over Norway, achieving state-of-the art performance in the simultaneous estimation of five complementary forest structure proxies. These two previous works were combined and expanded upon in Lang et al. (2023), where the authors proposed to train their CNN model ensemble using sparse 25 m footprint GEDI samples as ground-truth data. This resulted in the generation of a worldwide canopy height estimate map based only on Sentinel-2 acquisitions as input. The performance of the proposed approach was evaluated using a mixture of independent LVIS and ALS measurement campaigns, achieving an RMSE of 7.9 m and a mean error (ME) of 1.7 m.

To the best of our knowledge, the work in Carcereri et al. (2023) was the first one in the literature to investigate the use of a pure data-driven deep learning-based approach for forest height estimation from single-pass TanDEM-X InSAR data. The method consisted in a custom CNN architecture, trained on rasterized LVIS height estimates, acquired in the context of the 2016 AfriSAR campaign. This preliminary work achieved a MAE of 4.20 m, an RMSE of 5.69 m and a R^2 score of 0.73.

In light of the current state of the art, forest height estimation at large-scales using InSAR data is plagued by one or more operational compromises. Physical-based models, while offering a great mathematical interpretation behind the electromagnetic scattering mechanisms (Papathanassiou and Cloude, 2001), in practice require either a large-quantity of interferometric baselines, fully polarimetric InSAR acquisitions (Denbina et al., 2018), privileged ancillary information (e.g., DTM, LiDAR waveforms) (Guliaev et al., 2021; Choi et al., 2023), or simplifying assumptions that affect the inversion performance (Chen et al., 2016; Olesk et al., 2016). In practice, the estimation of the model parameters also requires extensive tuning with respect to the local properties of the forest: a requirement which is not suited for generalization purposes over larger areas. The overall performance has been shown to be outperformed by that achieved by data-driven approaches, even when the full model is inverted using privileged sources of information (Denbina et al., 2018; Lang et al., 2019; Carcereri et al., 2023). When it comes to state-of-the-art deep learning approaches, peak accuracy is currently achieved with optical data either by aggregating the estimates from multiple dates (Lang et al., 2019), or by means of model ensembles (Becker et al., 2023; Lang et al., 2023), both of which increase the computational complexity and the temporal delay between one estimate and the next. However, performance and operational deployment using optical sensors are especially limited by cloud coverage, with an estimated 50% of Earth's surface being hidden by clouds at any given moment (Gawlikowski et al., 2022). On the other hand, the few published works that explored the potential of SAR sensors (Becker et al., 2023) neither have considered the complexity given by the side-looking acquisition geometries, nor have explored the use of interferometric products, resulting in an accuracy which is worse than the one achieved with optical data only, and thus not justifying the added overhead and processing complexity.

In this work we present a robust deep learning approach which uses a single TanDEM-X bistatic acquisition to deliver state-of-the-art forest height estimates at large scale. Starting from the initial CNN architecture developed in Carcereri et al. (2023), we investigate the role of different input features and we design a novel training strategy tailored for an operational large-scale deployment. Finally, we combine the gathered information to generate a tree height map for the state of Gabon, obtained from estimates from a single TanDEM-X coverage (i.e., only one baseline) and subsequently moisaicked together. This makes our approach particularly interesting for the exploitation of the historical and current global TanDEM-X dataset (acquired since the end of 2010), as well as of the upcoming L-band NISAR mission (launch

planned in 2024, NASA/ISRO) and of the planned Sentinel-1 bistatic Earth Explorer mission Harmony (launch planned in 2029, ESA).

The paper is structured as follows. Section 2 lists the different datasets used in our approach. Section 3 starts with a brief introduction on the interferometric coherence. Then, it presents the details of our proposed deep learning approach, including the model architecture, the training strategy, the performance metrics and the developed approach to evaluate the trustworthiness of the model for large-scale inference. Section 4 illustrates a series of experiments for tacking the challenges of large-scale deployment, investigating the trade-offs between accuracy and model trustworthiness. This leads to the definition of the final model and to the generation of a country-scale map of canopy height over Gabon. Section 5 discusses our findings in the context of large-scale inference of forest parameters, highlighting potential issues and offering pragmatic solutions to model deployment and generalization capabilities. Finally, Section 6 summarizes our efforts and gives an outlook on potential future research aspects.

2. Materials

2.1. TanDEM-X data

The German TanDEM-X mission comprises two twin SAR satellites, TerraSAR-X and TanDEM-X, operating at X-band and flying in a varying close-orbit formation (Krieger et al., 2007; Zink et al., 2021). This particular configuration enables the acquisition of high-resolution, single-pass InSAR data with variable perpendicular baselines, allowing for the successful generation and delivery of a global digital elevation model (DEM) with unprecedented accuracy in 2016 (Rizzoli et al., 2017).

For the scope of this work, we considered TanDEM-X bistatic data, acquired in single polarization (HH) stripmap mode, with an extension in range of about 30 km. We distinguish two forms of TanDEM-X datasets:

- In order to properly train our model and generalize it across all possible acquisition geometries, we considered all existing TanDEM-X bistatic data acquired between December 2010 (i.e., the beginning of the mission) and 2022 over the five regions of interest (ROIs) of the 2016 AfriSAR campaign.
- For the generation of the final large-scale products, we retrieved all existing acquisitions covering the West Central African state of Gabon for the years of 2010/11. This allowed us to create one edge-to-edge mosaic, using products acquired with suitable interferometric baselines.

The resulting datasets are characterized by a large variety of acquisition geometries, i.e., of interferometric baselines and incidence angles. The inputs to our processing chain are the co-registered single-look complex (CoSSC) products. The underlying focusing and co-registration processing steps were performed by the operational TanDEM-X processor (ITP) (Fritz et al., 2012).

For each product we compute the backscattering coefficient σ^0, as recorded by the transmitting satellite only (monostatic channel). It is derived from the absolutely calibrated intensity β^0 (i.e., the radar brightness) and the local incidence angle θ_{inc} as:

$$\sigma^0 = \beta^0 \sin(\theta_{\text{inc}}), \tag{1}$$

where θ_{inc} is computed from the satellite's orbit position and the underlying DEM product.

For the estimation of the bistatic InSAR phase, we apply Φ-Net (Sica et al., 2021), a state-of-the-art residual deep-learning denoising architecture, capable of preserving the spatial resolution (compared to the commonly used boxcar multi-looking approach). For each input CoSSC product, we also generate the InSAR DEM, called raw DEM. For the sake of clarity, we recall that the value of an InSAR-based DEM, such as TanDEM-X, represents the topographic height corresponding to the location of the radar mean phase center. Given the capability of radar waves to penetrate into volumetric targets, such as vegetation, this elevation value is located somewhere below the top of the canopy, depending on the sensor characteristics (e.g., center frequency and acquisition geometry) as well as on the properties of the target itself. Differently, the terms digital surface model (DSM) and digital terrain model (DTM) identify the elevation of the top of the canopy and of the ground, respectively. The generation of the raw DEM is motivated by our previous conclusions in Carcereri et al. (2023) that the use of the global TanDEM-X edited DEM (González et al., 2020), generated by combining multiple acquisitions between 2010 and 2015, can be affected by small errors caused by the automatic editing procedure, which negatively impact our approach. By relying on the raw acquisition DEM we can also guarantee that all of our input features are consistent with each other. The raw DEM is also used to compute the local incidence angle θ_{inc} with respect to the local topography.

Additionally, by considering the annotated information on the satellites' positions, we encode the information about the interferometric acquisitions geometry in the form of a two-dimensional map of the height of ambiguity h_{amb}, which is defined as the vertical height change corresponding to a complete 2π phase cycle in the interferogram and it can be expressed for the single-pass InSAR case as:

$$h_{\text{amb}} = \frac{\lambda \cdot r \cdot \sin\theta_{\text{inc}}}{B_{\perp}}, \tag{2}$$

where $B_{\perp}$ is the orthogonal interferometric baseline, λ is the wavelength and r is the slant-range distance.

The interferometric coherence γ_{tot} represents the key metric to evaluate the interferometric performance, since it quantifies the amount of noise in the interferogram (Touzi et al., 1999). It is defined as the normalized cross-correlation coefficient of the interferometric image pair, called master (s_1) and slave (s_2), respectively:

$$\gamma_{\text{tot}} = \frac{E[s_1 \cdot s_2^*]}{\sqrt{E[s_1^2] \cdot E[s_2^2]}}, \tag{3}$$

where $E[\cdot]$ represents the expectation operator and $*$ the complex conjugate operator. As already done for the InSAR phase, also the interferometric coherence is estimated using Φ-Net (Sica et al., 2021). Following the approach presented in Rizzoli et al. (2022), it is possible to factorize γ_{tot} into its constituent error contributions, called decorrelation factors:

$$\gamma_{\text{tot}} = \gamma_{\text{SNR}} \cdot \gamma_{\text{quant}} \cdot \gamma_{\text{amb}} \cdot \gamma_{\text{rg}} \cdot \gamma_{\text{az}} \cdot \gamma_{\text{temp}} \cdot \gamma_{\text{vol}}, \tag{4}$$

where the different terms on the right-hand side identify the contributions due to limited signal-to-noise ratio (γ_{SNR}), quantization (γ_{quant}), ambiguities (γ_{amb}), baseline decorrelation (γ_{rg}), relative shift of the Doppler spectra (γ_{az}), temporal decorrelation (γ_{temp}) and volume decorrelation (γ_{vol}). In particular, the *volume decorrelation factor* γ_{vol} quantifies the degree of interferometric decorrelation caused by the scattering effects of a volumetric target, such as forests, sand or snow packs. From the total interferometric coherence we estimate γ_{vol}, by following the procedure presented in Rizzoli et al. (2022). The volume decorrelation factor constitutes a valuable proxy parameter for the vegetation structure as it is commonly modeled in the literature as the normalized Fourier transform of the vertical scatterer distribution (Papathanassiou and Cloude, 2001; Martone et al., 2016):

$$\tilde{\gamma}_{\text{vol}}(\vec{w}) = e^{ik_z z_0} \frac{\int_{z_0}^{z_0+h_v} F(z', \vec{w}) e^{ik_z z'} dz'}{\int_0^{h_v} F(z', \vec{w}) dz'}, \tag{5}$$

where $\vec{w}$ represents the polarization vector, $F(z', \vec{w})$ is the vertical scatterer distribution in the medium, z_0 is the ground elevation, h_v is the forest height and k_z is the vertical wavenumber. In turn, k_z is closely related to the height of ambiguity h_{amb} through $k_z = 2\pi/h_{\text{amb}}$ (Martone et al., 2016).

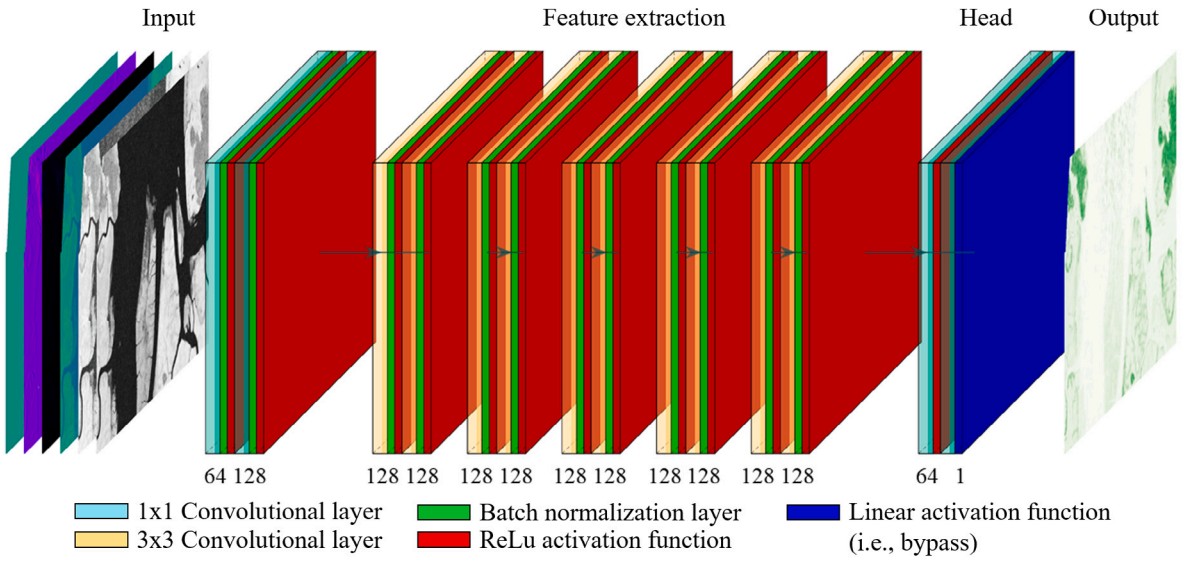


Fig. 1. The proposed fully convolutional deep learning model. The subscript numbers indicate the number of kernel filters.

2.2. AfriSAR-16 campaign

As the main source of reference forest height measurements we use the products generated from the 2016 AfriSAR campaign (Saatchi et al., 2019; Fatoyinbo et al., 2021) . We consider the full-waveform LiDAR measurements acquired by NASA's airborne LVIS (Land, Vegetation and Ice Sensor) instrument (Blair et al., 1999) between February and March of 2016. The laser shots sampled the ground in regular intervals, each of them covering a nominal footprint diameter of 18 m. The resulting vertical energy profiles were used to derive multiple forest structure proxy parameters, including the forest height estimates, expressed in terms of relative height (RH), which represents the height corresponding to a given percentile of returned energy. These products are also made available in the form of geomaps, which aggregate, interpolate and rasterize the discrete samples to generate a dense representation with a ground sampling distance of 25 m. For our work we use the rasterized RH99 statistic as our reference tree height measurement, since it represents a good proxy for the top of the canopy (99% of the returned energy), while it reduces the effects of strong outliers. We also use the rasterized DTM from LVIS to get the real topographic information below the dense canopy. The campaign measurements cover five heterogeneous areas within the state of Gabon (Saatchi et al., 2019):

- The Lopé National Park, consisting of a mixture of seasonal tropical forest and savannah, both affected by a distinct separation between wet and dry seasons (Guliaev et al., 2021). The area is characterized by strong topography, representing the highest elevation among all the considered regions of interest.
- The Mondah forest, which represents a small protected coastal site, partially flooded and characterized by the presence of both mangroves and tropical hardwood forests.
- The Mabounié site, which is a predominantly forested area, with localized sites of mostly anthropogenic degradation.
- The Pongara National Park, located on the southern side of the Gabon River and characterized by the presence of seasonally flooded forests, as well as very tall mangroves stands and some grassy savannah.
- The Rabi site, characterized by the presence of an onshore oil-drilling location, is largely covered by dense rainforest.

2.3. ESA WorldCover map 2021

In the pre-processing steps, we also make use of the *ESA 2021 WorldCover map*. This consists in a 10 m resolution global land-cover product that refers to 2021. It was generated using data from both ESA's Sentinel-1 and Sentinel-2 satellites and is freely accessible (Zanaga et al., 2022). We take advantage of the information it provides to mask out built-up areas and water bodies from our dataset.

3. Methods

In this section we present the details of the proposed approach, including the developed DL framework, the training and validation procedures, as well as the final inference step and reliability estimation.

3.1. Proposed deep learning framework

The proposed method relies on a deep learning architecture, in the form of a fully convolutional neural network (CNN). At its core, this technique consists of a sequence of linear cross-correlation computations, interleaved by non-linear operations (the so called activation functions). This typology of models has been at the heart of the AI surge in the computer vision field, as its major advantage over alternative architectures, such as fully-connected or transformer ones, is the computational efficiency in dealing with structured data including images. The working principle of CNNs exploits the typical spatial autocorrelation found in EO images. The convolution operation hard-codes this inductive bias by applying a small kernel function across the entire spatial extent of the input features, requiring only a limited amount of parameters in doing so. Assuming non-unitary kernels, by increasing the number of sequential cross-correlation calculations, the dimensions of the considered spatial contexts (the so-called receptive field) also increase. This in turn, allows for the creation of feature representations of increasing levels of abstraction and complexity. Crucially, this results in samples at the extremities of the receptive field being weighted less than those at its center.

Starting from the model proposed in Carcereri et al. (2023) and illustrated in Fig. 1, in this work we consider the following updated set of TanDEM-X-derived input features:

- The backscattering coefficient in HH polarization $\sigma^{0,\mathrm{dB}}_{HH}$ (in dB scale).
- The raw acquisition DEM h_{DEM}.

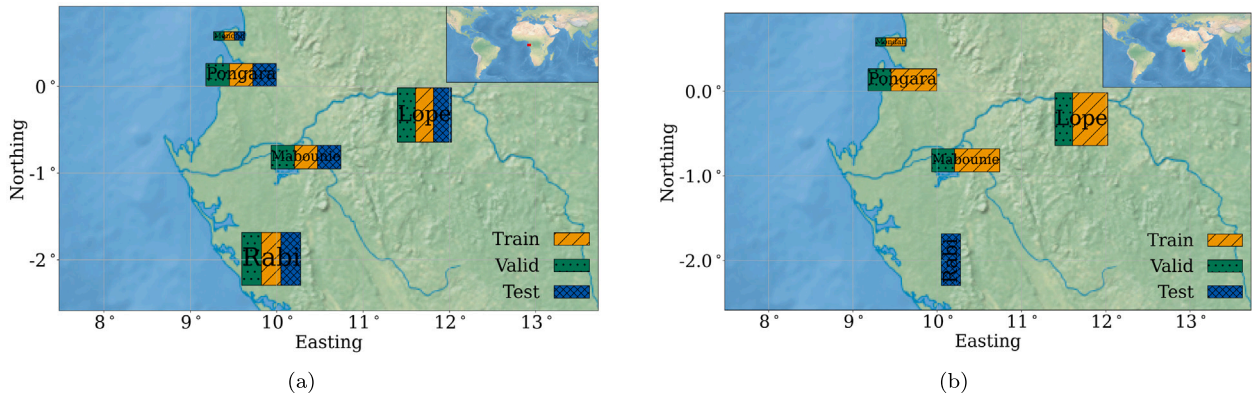


Fig. 2. (a) Geographic sub-setting of the AfriSAR campaign study areas into training, validation and testing, (b) Exemplary geographic sub-setting for the spatial cross-validation, considering Rabi a testing area only.

- The local incidence angle θ_{inc}.
- The estimated total interferometric coherence γ_{tot}.
- The estimated volume decorrelation factor γ_{vol}.
- The height of ambiguity h_{amb}.

As output we estimate the RH99 height metric as a proxy for the true top of the canopy.

The architecture can be split into several functional blocks. In the input block, the dimensionality of the input features is progressively increased first to 64 and then to 128 by means of two 1×1 convolution layers. This block is followed by a sequence of 5 hidden blocks, each consisting of two convolutional layers with 128, 3×3 kernel functions, and which can be interpreted as the main feature extraction sequence. In the output block (i.e., the tail of the model), the feature dimensionality is decreased back to 64 and later to a single output feature map by means of two additional convolution layers with 1×1 kernels. All convolution layers are followed by a rectified linear unit (ReLU) activation function and a batch normalization layer, except for the last one, which directly produces the output prediction.

3.2. Training, validation and testing strategy

In order to train, validate and test the proposed deep learning model, we split each of the five AfriSAR study areas into three equally sized sub-regions based on their geographic extent, and then associate these to either training, validation or testing, as presented in Fig. 2(a). This sub-setting strategy was chosen to guarantee the effective representation of the heterogeneous forests found across the study areas, while minimizing the effects of spatial autocorrelation-induced test bias that is commonly affecting random sampling strategies (Ploton et al., 2020; Kattenborn et al., 2022).

The model is trained using a fully supervised approach, consisting in the joint minimization of both the prediction error and l_2-norm of the model weights, and expressed by the following two term loss function:

$$Loss = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 + \lambda \cdot \sum_{j=1}^{m} w_j^2, \tag{6}$$

where $\hat{y}_i$ is the ith predicted sample, y_i is the corresponding ith reference sample value, w_j is the jth weight of the model, n is the total number of samples, m is the total number of weights and λ is the scaling factor of the l_2-norm.

The model's weights are iteratively updated on mini-batches of randomly sampled training patches. The size of these patches is 15×15 pixels and it has been chosen in accordance with the receptive field (RF) of our model, which, for a simple sequence of n two-dimensional convolutional layers with kernel size $k \times k$ pixels, can be computed as:

$$\text{RF} = n(k-1) + 1. \tag{7}$$

In our case, this results in a RF of 21 pixels or 525 m. Indeed, smaller patches would strongly crop the receptive field, while larger ones would not provide any additional benefit, coming at the cost of increased memory and computational loads, as well as poorer sampling of the available reference data.

Notably, the loss function is computed only on the central pixel of each patch, as this allows for better exploiting the available fragmented reference dataset and to provide a clearer interpretation of the model's working principle, as we will discuss in the inference post-processing. This choice limits each predicted center pixel to be seen only once for a given input image (i.e., no oversampling) acquired on a specific date and with a specific acquisition geometry (i.e., incidence angle and interferometric baseline). Furthermore, we allow pixels not covered by the reference data to be included inside the patches to give context to the forest boundaries.

During the backpropagation step we make use of the commonly employed ADAM optimizer (Kingma and Ba, 2017). We use the default hyperparameters of the Keras implementation, except for the initial learning rate, which is set to 10^{-4}. We determine the end of the training phase on the dedicated validation set by applying an early-stopping criterion once the model has stopped improving for more than 35 consecutive epochs.

A total of $13 \cdot 10^6$ patches is available for training, which on a single NVIDIA A100 GPU takes a maximum of 9 hours to train following the described strategy.

In order to test the trained model, we apply the inference directly at image level, by splitting each image into smaller chunks of data (2000×2000 pixel) to fit the GPU's VRAM buffer requirements. To provide enough contextual information to the model, we mask out the inferred pixels which do not correspond to a full valid neighborhood equal to the training patch-size. By applying this condition to all border pixels, we effectively delete missing values inside the image, requiring the inference chunks to be sampled with overlap in order to reconstruct a contiguous prediction map.

We test our predictions by comparing them on a pixel-wise level with the corresponding values in the reference data. To evaluate the performance of our model, we use the mean error (ME), the mean absolute error (MAE), the mean absolute error (MAPE), the root mean squared error (RMSE) and the coefficient of determination (R^2), defined as follows:

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right), \tag{8}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|, \tag{9}$$

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|, \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}, \tag{11}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}, \tag{12}$$

where $\bar{y}$ it the mean reference value.

Furthermore, to test the capability of the model to generalize over unseen regions and to assess the possible spatial correlations between the areas used for training and testing (caused by the vicinity of the split subsets in each AfriSAR test site), we perform geographic cross-validations by iteratively excluding one of the five sites from training and validation. To compensate for the decrease in training samples, we assign $\frac{2}{3}$ of each of the remaining four sites for training. An example is depicted in Fig. 2(b) for the permutation in which the site of Rabi is used for testing only. The concept is then repeated for all different permutations of the five AfriSAR test sites.

3.3. Country-scale inference and Map of Applicability (MoA)

For the generation of the final country-scale CHM mosaic we consider TanDEM-X acquisitions from the first global coverage, designed for the generation of the global DEM product (Rizzoli et al., 2017). In this way, we can guarantee an almost complete coverage with minimal gaps, as the TanDEM-X products are otherwise acquired irregularly and depending on the specific acquisition planning. Conversely, some regions are imaged multiple times per year (e.g., the AfriSAR test sites) and overlaps are therefore still possible. In order to evaluate the single-baseline quality of the proposed method, we first compute the mean h_{amb} of the overall distribution seen during training. Then, in presence of overlapping acquisitions only the one with the minimum distance from this value is considered. As described in Section 3.1, the model is subsequently applied to each acquisition individually to generate its corresponding CH estimate. Finally, the resulting list of forest height maps is mosaicked together to generate the country-scale product.

The independent validation of the product is challenging: no field plots exist at the considered scale and they are typically limited to the areas covered by the AfriSAR campaign, while spacerborne LiDAR missions such as the Global Ecosystem Dynamics Investigation (GEDI) (Dubayah et al., 2020) are reportedly ill-suited as sources of reference data in the presence of tall and dense canopy (Fayad et al., 2022; Lahssini et al., 2022; Morin et al., 2022). Inspired by the work in Meyer and Pebesma (2021), we propose to assess the reliability of the model's predictions by validating that the input predictors fall within the subspace sampled by the training data, as data-driven methods fail to perform reliably on out-of-distribution (OOD) predictor combinations (Liu et al., 2020). Ideally, such an evaluation would be performed by exhaustively determining for each inference sample the minimum distance in the predictor hyperspace to the training data and determining, on the basis of test data, where the trustworthiness of the model falls off. Unfortunately, such a computation becomes computationally intractable for large numbers of training or inference samples such as those considered in this study.

To overcome this issue, we instead propose the definition of an ad-hoc ODD detector relying on an approximation of the joint predictor distribution as a proxy for the training set-sampled predictor space. To obtain such an estimate, we start by computing the histograms for the individual predictors across globally-defined value ranges.[1] It is then possible to compute the relative frequency (i.e. the density) value $d_{j,i}$ for the ith-bin and the jth-predictor as:

$$d_{j,i} = 100\% \cdot \frac{h_{j,i}}{\sum_{n=1}^{N} h_{j,n}}, \tag{13}$$

where $h_{j,i}$ is the absolute frequency value for the ith-bin and the jth-predictor, and N is the total number histogram bins. At inference, each predictor is associated to the density value $d_{j,i}$ of the corresponding ith-bin. This leads to the generation of a geographic map representing, for each pixel position (x, y), the relative sample frequency $d_{j,x,y}$ seen by the model during training. The individual predictor maps are then aggregated into a single reliability score map $S_{x,y}$ by computing their geometric mean as:

$$S_{x,y} = \left(\prod_{j=1}^{J} d_{j,x,y}\right)^{\frac{1}{J}}, \tag{14}$$

where J is the number of predictors. The resulting mean density map is directly correlated to the model reliability, as values at (or close to) zero have jointly seen no (or few) samples in the corresponding predictor sub-space. Finally, the validation set is used to determine the threshold for $S_{x,y}$ that minimizes the prediction MSE, allowing for the generation of a binary Map of Applicability (MoA). Values below such a threshold are considered unreliable and can therefore be discarded.

In practice, the training set predictor distribution and the threshold are pre-computed once for each independent model and used at inference to estimate the areas of low prediction reliability. By combining the performance metrics introduced in Section 3.2 with the proposed MoA we drive the joint definition of a proper training dataset and of a set of predictors, which can yield the best possible trade-off between tested performance and model trustworthiness for large-scale inference. This is the reasoning behind the series of experiments proposed in Section 4.

4. Experiments and results

4.1. The impact of missing predictor representation

The first experiment that we propose considers the application of the training, validation and testing strategy presented in Section 3.2. In order to be as consistent as possible with the LVIS reference dataset, we select TanDEM-X data acquired during 2015–2016 only. The performance of the resulting baseline model is summarized in Table 1 (a) for each test site separately and overall, achieving a ME of −0.96 m, a MAE of 4.05 m, a MAPE of 14.28%, an RMSE of 5.31 m and a R^2 of 0.75. Moreover, we also perform a cross-validation test, as presented in Section 3.2, and the results are summarized in Table 1 (b). The performance metrics confirm that the model is robust also when tested on totally independent test sites. Only a small loss in performance with respect to the baseline case is detected (ME of −0.52 m, MAE of 4.54 m, MAPE of 16.38%, RMSE of 5.94 m and R^2 of 0.69). This is limited to the Lopé and Pongara test permutations, and can be explained by the unique phenological and topographical characteristic found in these sites.

We then apply the derived baseline model at country-scale by considering all available TanDEM-X acquisitions acquired between December 2010 and the end of 2011 in correspondence of the first mission global coverage. We generate a large-scale CHM mosaic and the corresponding reliability score map and MoA as presented in Section 3.3. The results are depicted in Fig. 3. As it can be seen, the MoA presents extended regions of zero values which correspond to entire TanDEM-X data-takes, revealing missing representations mainly associated to TanDEM-X acquisition-related parameters. When analyzing in depth the actual contribution of each single predictor to the reliability score map, as presented in Fig. 4, one can note that the most critical predictor is the height of ambiguity, which is directly related to the InSAR

[1] For each predictor, we consider the following ranges: $\sigma_{HH}^{0,\mathrm{dB}} \in [-25, 10]$, $\theta_{\mathrm{inc}} \in [0, -\frac{\pi}{2}]$ rad, $\gamma_{\mathrm{tot}} \in [0, 1]$, $\gamma_{\mathrm{vol}} \in [0, 1]$, $h_{\mathrm{amb}} \in [15, 120]$ m, $h_{\mathrm{DEM}} \in [0, 1100]$ m, $h_{\mathrm{DEM}}^{\mathrm{HPF}} \in [-700, 700]$ m, $\nabla h_{\mathrm{DEM}} \in [-4, 4]$ m m^{-1}, where the last two predictors are introduced later on in Section 4.3.

Table 1
(a) Performance metrics computed for the model trained using TanDEM-X data acquired in 2015–2016 only, for each AfriSAR test site, separately, and overall. (b) Performance metrics computed for the cross-validation experiment, for each AfriSAR test site permutation, separately, and overall.

(a)

2015–2016 dataset performance

Experiment	ME [m]	MAE [m]	MAPE [%]	RMSE [m]	R^2 [·]
Lope	−0.42	4.12	11.25	5.34	0.40
Mabounie	−1.03	4.82	15.84	6.23	0.37
Mondah	0.89	2.25	28.57	3.22	0.90
Pongara	−0.05	3.03	17.30	4.28	0.92
Rabi	−2.03	4.18	13.58	5.32	0.53
Overall	−0.96	4.05	14.28	5.31	0.75

(b)

Cross-validation performance

Experiment	ME [m]	MAE [m]	MAPE [%]	RMSE [m]	R^2 [·]
Lope	−0.24	5.23	15.27	6.75	0.04
Mabounie	−0.25	4.79	16.20	6.21	0.38
Mondah	0.84	2.35	29.50	3.32	0.90
Pongara	−1.60	4.56	23.27	5.97	0.84
Rabi	−0.79	4.03	13.54	5.18	0.56
Overall	−0.52	4.54	16.38	5.94	0.69

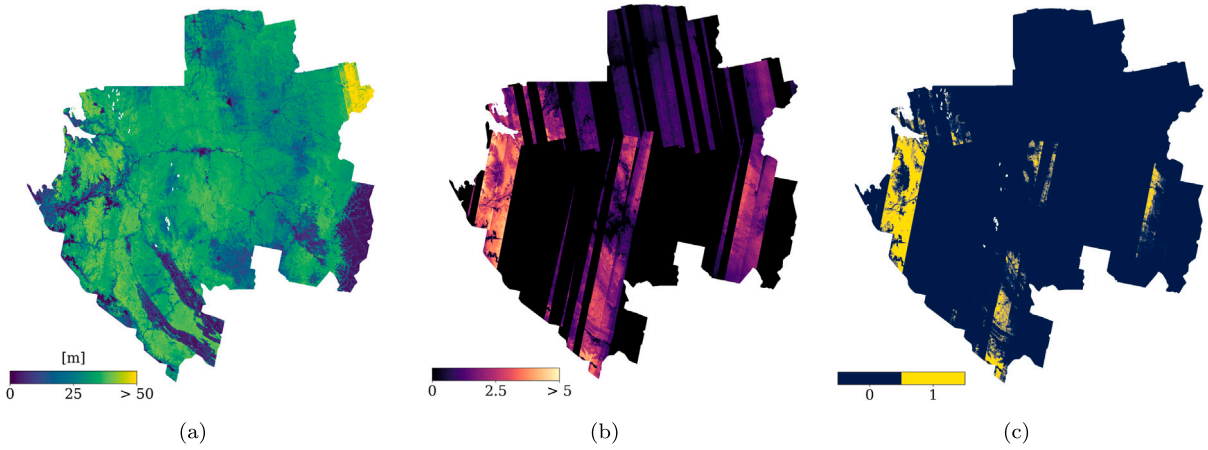


Fig. 3. (a) Country-scale CHM generated using the baseline model trained using TanDEM-X from 2015–2016 only, (b) Corresponding reliability score map and (c) binary MoA.

acquisition geometry. Nevertheless, also the raw DEM relative sample frequency map presents extended areas of zero values, significantly contributing to the unreliability of the country-scale CHM mosaic. This is reflected in the histograms of such features for the training and inference datasets, respectively, as presented in Fig. 5. Consistently, severe underestimation of the CHM can be seen in the country-scale mosaic in Fig. 3(a) in correspondence of zero values of the MoA. Therefore, at the present stage, the model cannot be considered to be reliable outside of the AfriSAR test regions. Possible solutions to solve these issues are proposed in Sections 4.2 and 4.3, respectively.

4.2. Height of ambiguity analysis

In order to tackle the challenge of missing representation in the input heights of ambiguity, we propose a relaxation of the temporal stationarity constraint between the 2016 LVIS reference measurements and the input TanDEM-X data considered in Section 4.1, where only data-takes acquired in 2015/2016 were considered. This results in a larger compatibility with the existing TanDEM-X archived data and leads to a more representative selection of acquisition geometries.

We achieve this by considering TanDEM-X data covering the AfriSAR test sites, acquired over a time span of about 11 years, starting from the end of 2010 (beginning of the bistatic TanDEM-X mission) up to 2021. This allows for the generation of a complete dataset characterized by the distribution of h_{amb} presented in Fig. 6. As it can be seen, the distribution of the h_{amb} used for the country-scale inference depicted in Fig. 5(b) is much better represented with respect to the initial 2015/2016 case. By considering multiple acquisitions we also allow our model to learn a more robust relationship between the acquisition conditions of the input imagery and the reference canopy height. This assumes that the temporal misalignment between the input and the reference data results only in minor forest height inconsistencies due to natural phenomena, such as growth and tree replacement. These can be characterized as an additive noise contribution, and can thus be interpreted as a data augmentation process. On the other hand, drastic logging, fire or afforestation events, if present, are assumed to be limited in scope and are considered as outliers, whose effects are mitigated by the availability of a large number of training samples.

Regarding the model performance, we test only on TanDEM-X data acquired between 2015 and 2016, in order to be as consistent as possible with the reference LVIS data, as well as with the settings of the previous experiment in Section 4.1. The achieved model performance is summarized in Table 2 (a) (ME of −0.54 m, MAE of 3.78 m, MAPE of 13.08%, RMSE of 4.98 m and R^2 of 0.78), which shows an overall improvement with respect to the initial baseline scenario. The cross-validation results are summarized in Table 2 (b), confirming the robustness of the model when tested on completely independent areas.

The resulting country-scale CHM mosaic, generated from TanDEM-X acquisitions from December 2010 up to the end of 2011 (as done in Section 4.1), the corresponding reliability score map and MoA are depicted in Fig. 7. The MoA presents now much less zero values with respect to the initial model presented in Fig. 3(b), and the remaining critical areas are not primarily linked to the TanDEM-X acquisition geometry but rather to local topographic effects only. Moreover, the areas of severe CHM underestimation shown in the CHM mosaic of Fig. 3(a) are not present anymore in the new mosaic of Fig. 7(a). From now on we will therefore only consider the model trained using the extended TanDEM-X dataset for further experiments on the impact of missing representations of the DEM predictor in Section 4.3.

4.3. DEM analysis

When considering the issue of missing DEM representations in the training set, it is not possible to follow the same strategy proposed for the height of ambiguity in Section 4.2, since the inclusions of new acquisitions over the same initial regions of interest would not allow

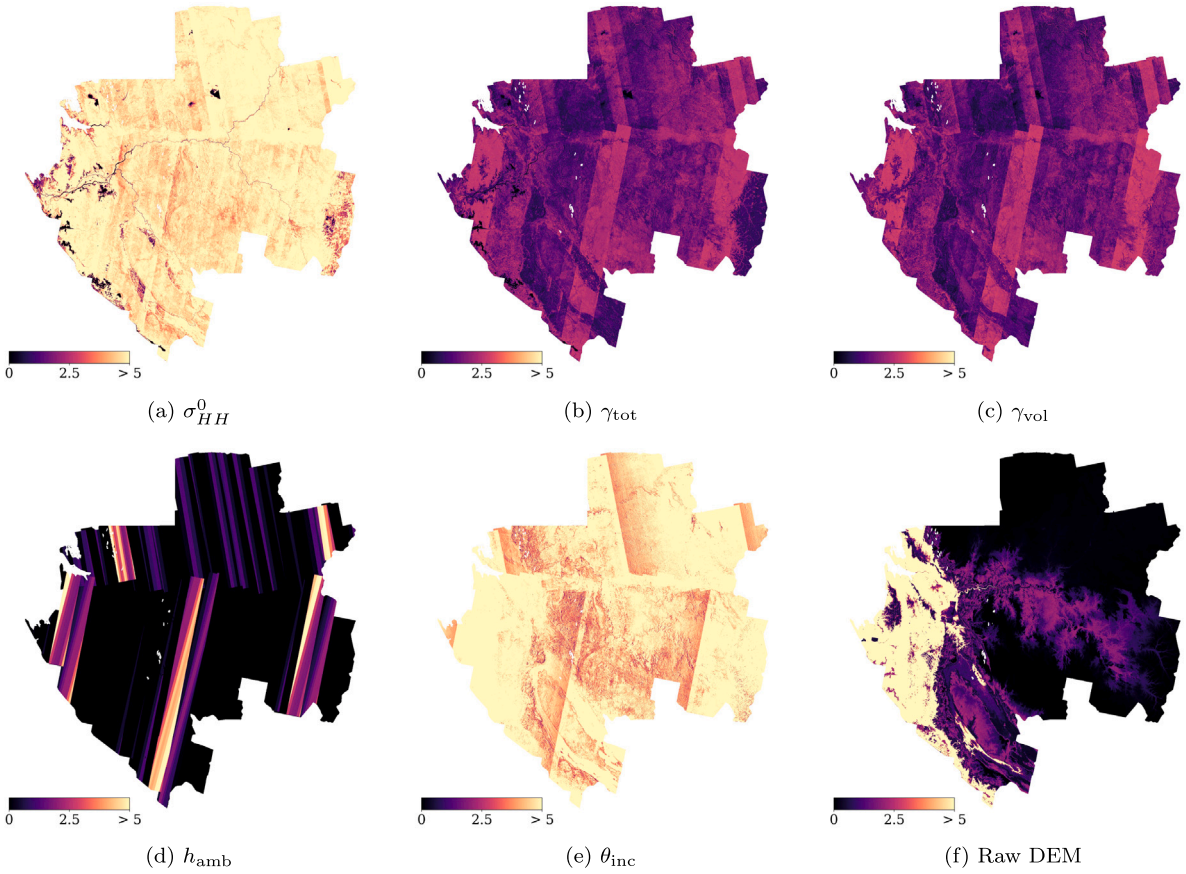


Fig. 4. Relative sample frequency for each predictor (indicated below each image) of the model presented in Section 4.1, trained with TanDEM-X data acquired in 2015/2016 only.

Table 2
(a) Performance metrics computed for the model trained using the extended set of TanDEM-X data acquired in between 2010 and 2021, for each AfriSAR test site, separately, and overall. (b) Performance metrics computed for the corresponding cross-validation experiment, for each AfriSAR test site permutation, separately, and overall.

(a)

Extended dataset performance

Experiment	ME [m]	MAE [m]	MAPE [%]	RMSE [m]	R^2 [·]
Lope	−0.01	3.88	10.68	5.02	0.47
Mabounie	0.02	4.56	15.49	5.91	0.44
Mondah	−0.15	2.15	22.81	3.12	0.91
Pongara	−0.90	2.84	14.91	4.00	0.93
Rabi	−1.29	3.81	12.36	4.93	0.60
Overall	−0.54	3.78	13.08	4.98	0.78

(b)

Cross-validation performance

Experiment	ME [m]	MAE [m]	MAPE [%]	RMSE [m]	R^2 [·]
Lope	−0.22	4.29	12.15	5.52	0.36
Mabounie	1.58	4.86	17.19	6.31	0.36
Mondah	0.87	2.19	26.26	3.17	0.91
Pongara	−1.38	4.95	26.64	6.37	0.81
Rabi	−0.80	3.84	12.80	4.97	0.59
Overall	−0.18	4.20	15.43	5.49	0.73

the model to see a larger variety of topographies during training. In particular, when comparing the histograms in Fig. 5(c) and (d), one can note that only elevations up to about 400 m are well represented by the AfriSAR test sites. At inference, this results in the majority of the elevation samples being poorly or not at all represented during the training phase.

To address this issue, we propose to either completely remove the DEM as a predictor or to substitute it with some proxy variables which describe only local topographic variations instead of the absolute elevation of the scene. Regarding the former solution, we expect to loose some performance with respect to the results presented in Section 4.2 in favor of a more robust model, while, regarding the latter, we investigate the use of two different DEM-derived variables: the estimates for the set of spatial partial derivatives ∇h_{DEM}, which correspond to the estimation of the local terrain slope, and a high-pass filtered version of the DEM $h^{\text{HPF}}_{\text{DEM}}$, which removes the mean elevation of the scene, highlighting only local high-frequency variations of the topography. One should note that ∇h_{DEM} is computed as:

$$\nabla h_{\text{DEM}}(x, y) = \left(\frac{\partial h_{\text{DEM}}(x, y)}{\partial x}, \frac{\partial h_{\text{DEM}}(x, y)}{\partial y} \right), \tag{15}$$

where x and y are the horizontal and vertical coordinates, respectively. This corresponds to the addition of two different input predictors, identifying the horizontal and vertical partial derivatives, respectively. The performance for all different test cases is summarized in Table 3, together with the performance of the model derived from the extended dataset presented in Section 4.2 for comparison purposes (**Baseline** case). As expected, the complete removal of the DEM from the set of predictors (**w/o DEM** case) causes a general loss in performance, which is partly mitigated by the use of the DEM spatial derivatives (**∇** case) or the high-pass filtered version (**HPF** case), with the former achieving the overall peak performance (ME of 0.12 m, MAE of 3.90 m, MAPE of 15.38%, RMSE of 5.08 m and R^2 of 0.77). On the other hand, the robustness of model for country-scale inference significantly improves, as it can be seen from the MoA of all three considered cases in Fig. 8.

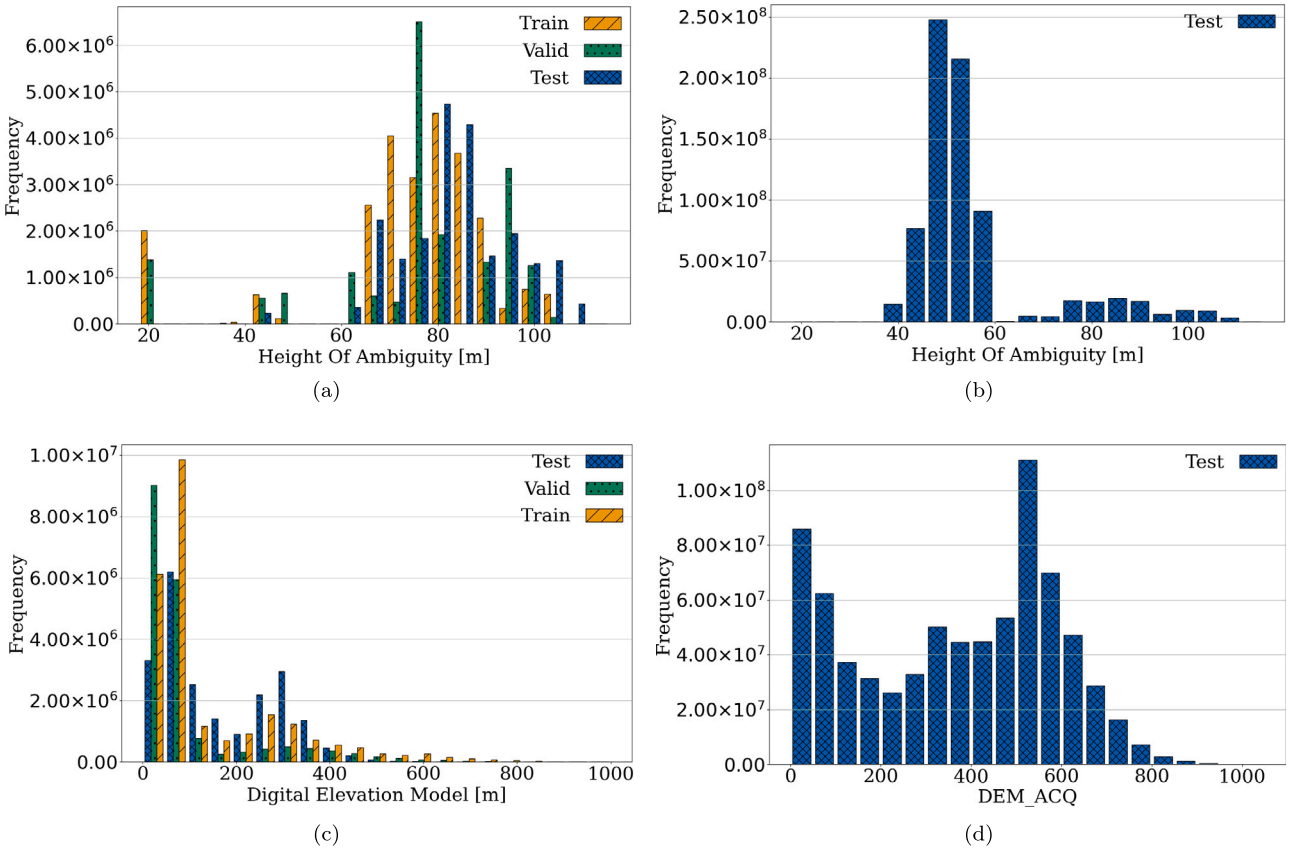


Fig. 5. (a) Height of ambiguity histogram for the training, validation and testing datasets corresponding to the 2015–2016 TanDEM-X acquisitions, (b) height of ambiguity histogram for the 2010–2011 TanDEM-X acquisitions used for the country-scale inference. (c) DEM histogram for the training, validation and testing datasets corresponding to the 2015–2016 TanDEM-X acquisitions, (d) DEM histogram for the 2010–2011 TanDEM-X acquisitions used for the country-scale inference.

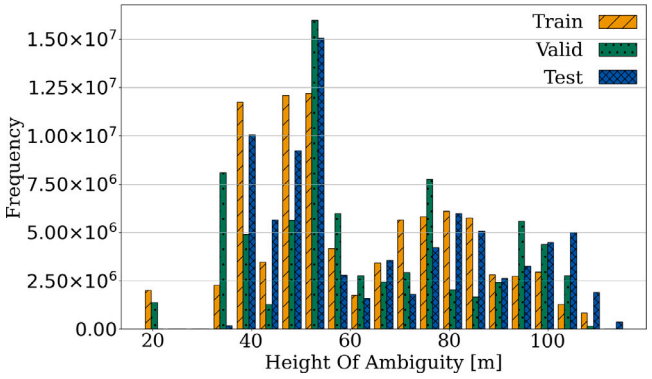


Fig. 6. Hight of ambiguity histogram for the extended training, validation and testing dataset, comprising TanDEM-X acquisitions covering the AfriSAR test sites from December 2010 up to the end of 2021.

Indeed, for all considered solutions the corresponding MoAs show an almost complete coverage of valid values, with the **∇** case (subfig. (b) and (e)) being characterized by overall higher values in the reliability score map. The majority of remaining invalid values is located in correspondence of water surfaces, which were not seen during training.

4.3.1. Final country-scale model and inference

In light of the knowledge gained from the previous experiments, we define our final model for the generation of a country-scale map of the canopy height over Gabon as the CNN architecture proposed in Section 3.1, trained with an extended dataset of TanDEM-X image acquired from December 2010 up to 2021 over the test sites of the 2016 AfriSAR campaign as presented in Section 4.2, and replacing the input DEM predictor with the estimate of its spatial derivatives, as proposed in Section 4.3.

Table 3
Performance metrics for the DEM analysis experiment. Each row identifies the performance of the model derived from the extended dataset presented in Section 4.2 (**Baseline**), the model trained without the DEM as predictor (**w/o DEM**), the model trained with the DEM derivatives (**∇**) and the model trained with the high-pass filtered version of the DEM (**HPF**).

DEM analysis performance					
Experiment	ME [m]	MAE [m]	MAPE [%]	RMSE [m]	R^2 [·]
Baseline	−0.54	3.78	13.08	4.98	0.78
w/o DEM	−0.58	4.13	15.93	5.35	0.75
∇	0.12	3.90	15.38	5.08	0.77
HPF	0.11	4.00	15.26	5.23	0.76

The detailed performance metrics are reported in Table 4 (a) (ME of 0.12 m, MAE of 3.90 m, MAPE of 15.38%, RMSE of 5.08 m and R^2 of 0.77). Compared to the reference model presented in Section 4.2 only a minor degradation in performance can be observed, predominantly caused by a small overestimation of very short vegetation samples. This behavior can be spotted in the scatterplot presented in Fig. 9 (a), showing the reference RH99 LVIS values versus the prediction for the final selected model. The estimation bias (ME) with respect to different reference tree height sub-ranges is shown in 9 (b), together with the overall reference RH99 distribution of test samples. Notably, measurements are on average slightly overestimated for vegetation heights below 15 m, are unbiased between 30 m and 40 m, with a tendency to more strongly overestimate forest heights in the 15 m to 30 m range and to underestimate for values above 40 m. The results

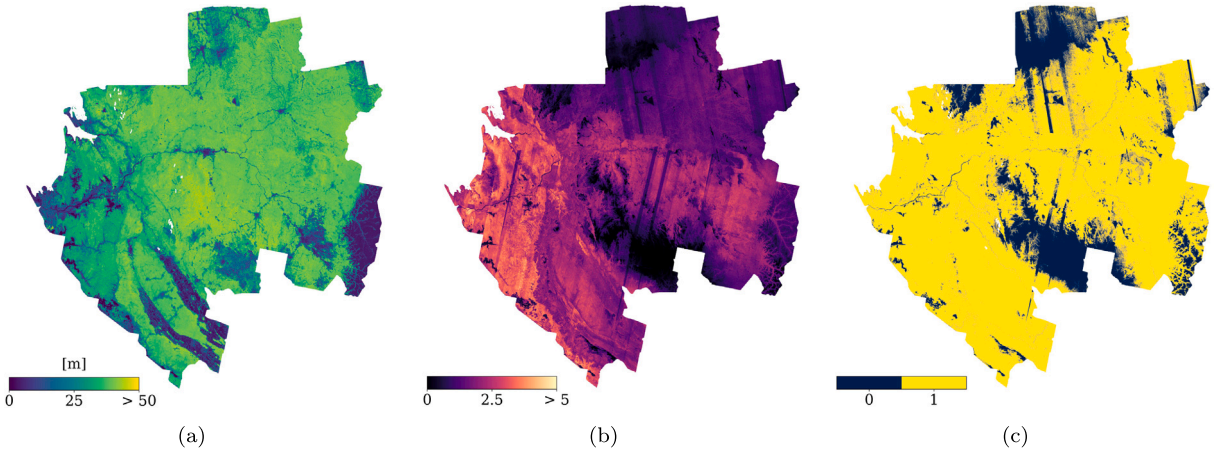


Fig. 7. (a) Country-scale CHM generated with the model trained using the extended dataset comprising TanDEM-X from 2010 up to 2021 (Section 4.2), (b) corresponding reliability score map and (c) binary MoA.

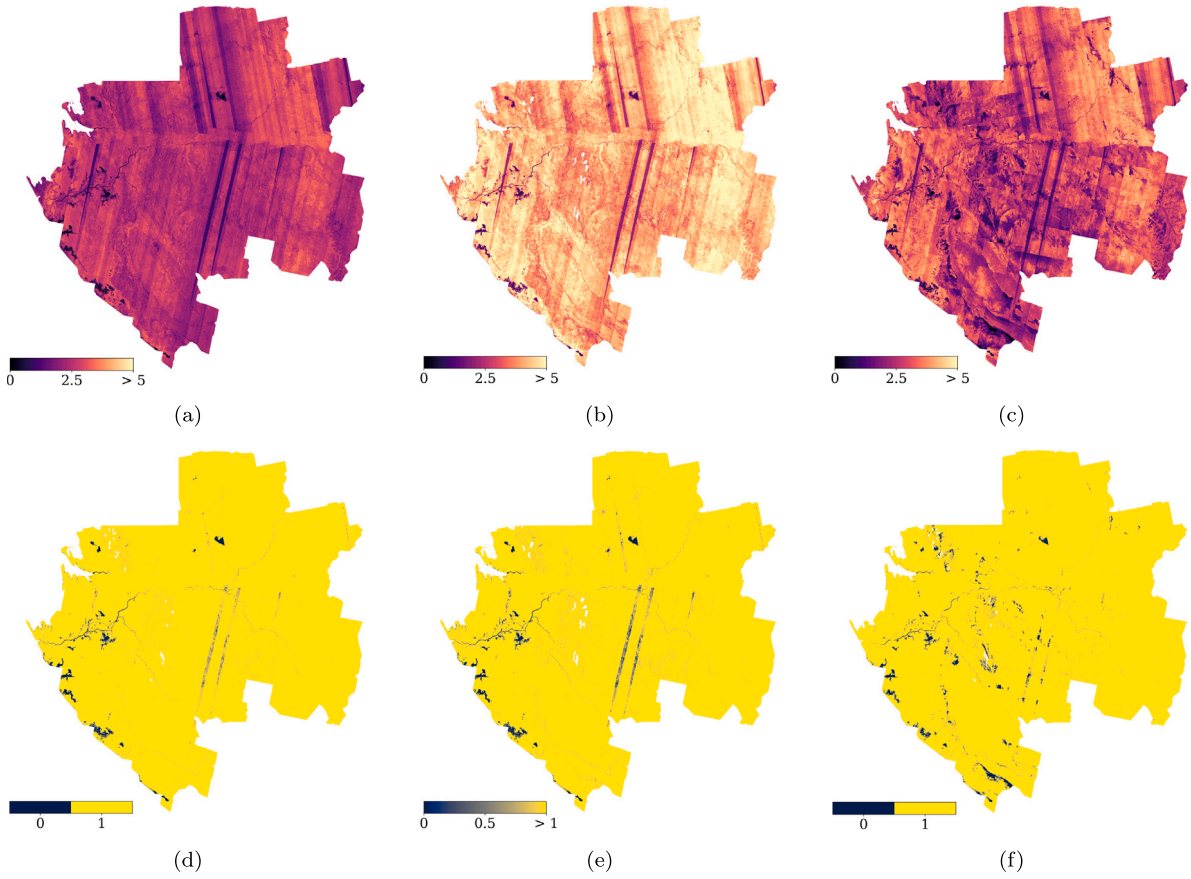


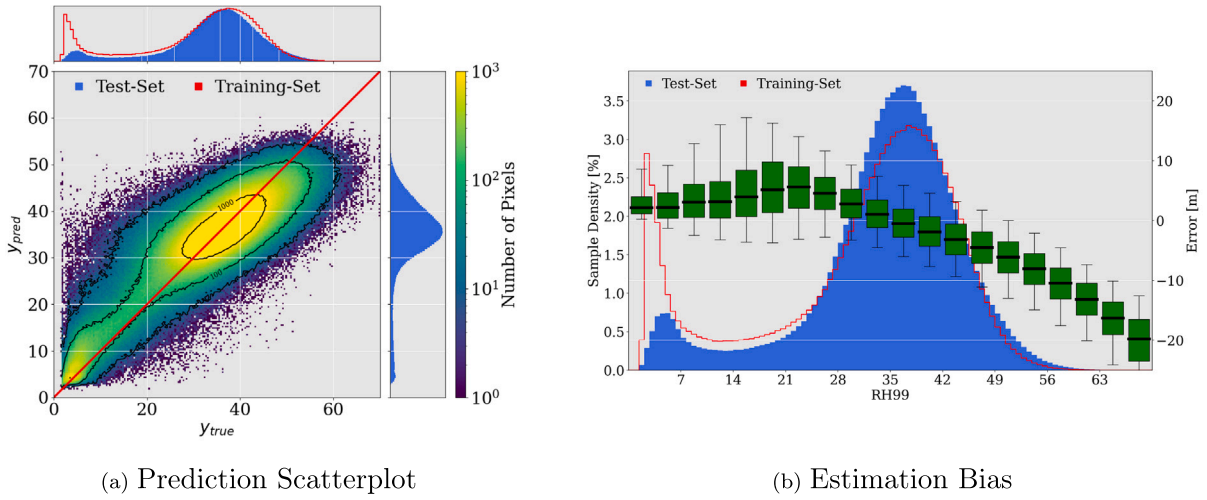
Fig. 8. (Top row) Reliability score maps for the three different DEM-related solutions: (a) removal of the DEM as predictor, (b) substitution of the DEM with its spatial derivatives (local slope), (c) substitution of the DEM with its high-pass filtered version. (Bottom row) Corresponding MoAs.

for the geographic cross-validation experiments are reported in Table 4 (b) and show a similar degree of spatial independence as seen for the previous results.

To further analyze the model's bias with respect to parameters characterizing both the illuminated areas and the acquisition geometry, we extend the test dataset to all TanDEM-X acquisitions from 2010 up to 2021 overlapping the AfriSAR test sites. In this way, a good representation of the analyzed parameters is considered. Fig. 10(a) displays the relative dependency of the estimation error on the acquisition Day Of the Year (DOY), highlighting a comparably unbiased relationship across the value range. Fig. 10(b) relates the error to the local terrain slope, which is computed using the LVIS-derived DTM estimates. It is possible to note that the median absolute error show almost no dependency on the local slope of the underlying topography. Additionally, we evaluate the performance dependency on the TanDEM-X acquisition geometry. We observe that across both the h_{amb} (Fig. 10(c)) and the incidence

Table 4
(a) Performance metrics computed for the final model trained using the estimate of the spatial DEM gradient as a replacement for the DEM itself, shown for each AfriSAR test site separately and overall. (b) Performance metrics computed for the corresponding cross-validation experiment, shown for each AfriSAR test site permutation separately and overall.

∇ performance						Cross-validation performance					
Experiment	ME [m]	MAE [m]	MAPE [%]	RMSE [m]	R^2 [·]	Experiment	ME [m]	MAE [m]	MAPE [%]	RMSE [m]	R^2 [·]
Lope	−0.02	3.78	10.41	4.88	0.50	Lope	−3.82	5.27	13.36	6.65	0.07
Mabounie	0.44	4.45	15.24	5.77	0.46	Mabounie	1.50	4.65	16.19	6.07	0.41
Mondah	1.98	2.95	39.69	4.00	0.85	Mondah	3.25	4.12	62.94	5.71	0.69
Pongara	1.52	4.21	28.27	5.45	0.86	Pongara	−1.17	5.37	28.91	7.01	0.77
Rabi	−0.59	3.79	13.01	4.93	0.60	Rabi	−1.18	3.85	12.84	4.96	0.59
Overall	0.12	3.90	15.38	5.08	0.77	Overall	−1.43	4.64	17.69	6.03	0.68



(a) Prediction Scatterplot (b) Estimation Bias

Fig. 9. Comparative estimation performance plots between the selected model for country-scale inference and the reference LVIS dataset. The scatterplot (a) displays the linear prediction agreement. The boxplot sequence (b) captures the estimation bias and spread for different reference tree height sub-ranges; the whisker contain 90% of the samples, the boxes 50%, while the black line represents the median value. The background histograms depict the relative samples distributions of the training (red) and test sets (blue), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

angle ranges (Fig. 10(d)) the median estimation errors suggest an essentially unbiased estimation.

Finally, the country-scale mosaic of the CHM over Gabon at 25 m resolution, inferred for the TanDEM-X acquisitions of 2010/2011 is depicted in Fig. 11(a). Further details at higher resolution, corresponding to the areas included in the red squares, are presented in Fig. 11(b). The first zoom-in (left), shows the presence of primary tropical forest, a complex system of rivers and antropogenic activities. In the second one (center) it is possible to observe the presence of tall mangroves along the shores of the Gabon estuary, with peak canopy heights above 45 m. The third one (right) shows a further example of a dense mangrove forest along the coast.

5. Discussion

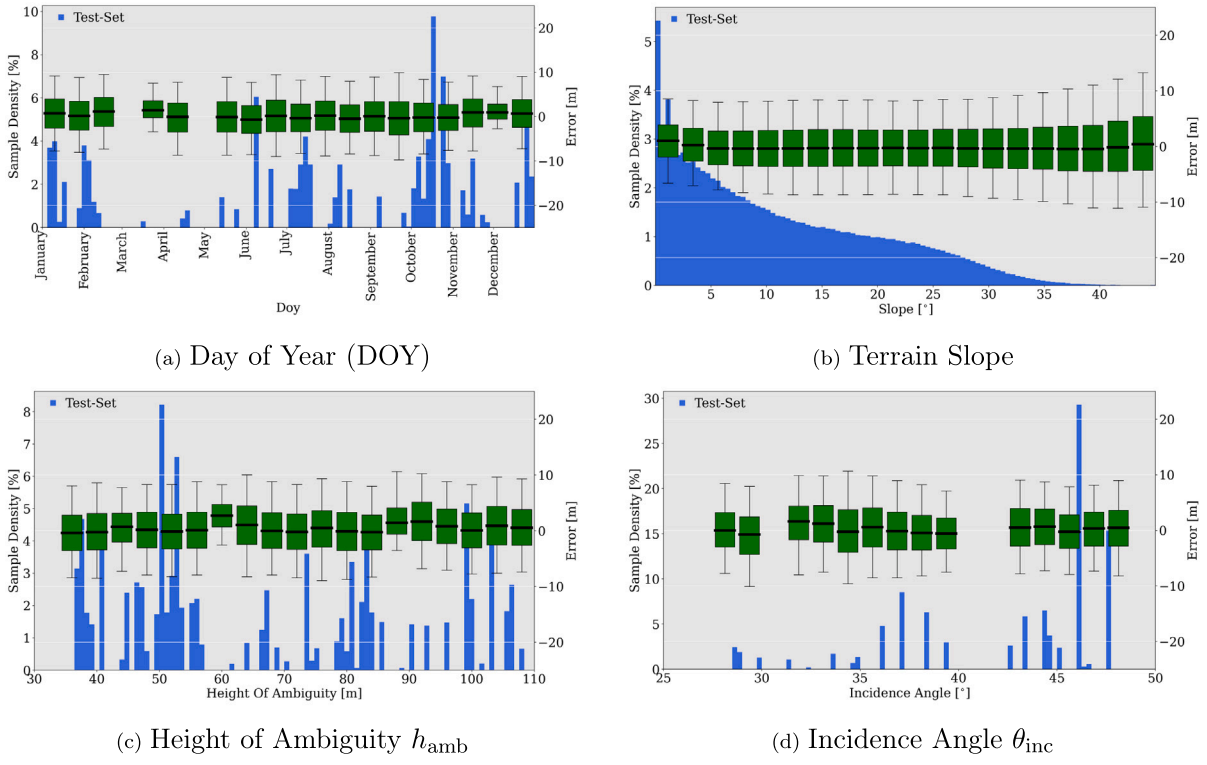
The results detailed in Section 4 provide a complete overview on the challenges and solutions related to the country-scale application of the proposed method. In particular, moving from the confined study areas of the 2016 AfriSAR campaign to the country-scale inference poses the natural challenge of validating the final product, especially in the absence of complementary reference measurements.

We approached the problem following two different paths. On the one hand, we examined the reliability of the achieved performance metrics by verifying the absence of spatial correlation between the training and test sets. To do so, we carried out geographically-independent cross-validations, which resulted in consistently small deviations in performance with respect to the standard testing strategy. In particular, these are limited to the Lopé and Pongara test permutations, being the former characterized by high-relief terrain, and the latter by the presence of tall mangroves. These observations allow us to confirm the soundness of the achieved model performance.

On the other hand, we assess the applicability of the model at country scale, where the estimates cannot be validated otherwise. Based on the assumption of a unique bijective relationship between predictors and forest height, the proposed approach is meant to identify those predictors positioned inside the subspace sampled by the training dataset. Rather than providing a pixel-wise validation, this approach allows for assessing the trustworthiness of the CH estimates. This means that we are not pixel-wise associating an accuracy value to each estimate, but are instead able to identify whether the model accuracy falls within the boundaries defined during the test phase. Clearly, should the underlying assumption of a bijective relationship not hold anymore (i.e., by missing a necessary discriminative feature), also the MoA would fail to detect unreliable estimates.

In practice, in the analyses in Sections 4.1 and 4.2 the MoAs have allowed for detecting missing representations in the training data, which directly match with strongly underestimated forest heights. On the contrary, the proposed modifications of the training data set in Section 4.2 and the new input predictors defined in Section 4.3 have led to the definition of a robust model for country-scale inference.

The final model achieves an overall very competitive performance, which starts to be significantly biased towards underestimation only for canopy heights above 45 m. Moreover, beyond canopy heights of 55 m, the model tends to saturate, a behavior often observed in the literature (Lang et al., 2019; Wagner et al., 2024; Schwartz et al., 2024). This effect might be related to different aspects. On the one hand, it could be related to the disproportionately low frequency of high tree samples in the training set or, on the other hand, to the limited capability of radar waves at X-band to penetrate into dense forests.



(a) Day of Year (DOY)

(b) Terrain Slope

(c) Height of Ambiguity h_{amb}

(d) Incidence Angle θ_{inc}

Fig. 10. Estimation error bias and spread versus the acquisition DOY (a), the Terrain Slope (b), the h_{amb} (c) and the incidence angle (d) features, represented as a discrete sequences of boxplots. Each boxplot covers a feature sub-range, and is described by its whiskers (containing 90 percent of the samples), its box (containing 50 percent of the samples) and black median line. In the backgrounds, the respective feature distributions on the test set (in blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Regarding the fact that the performance remains stable with respect to the DOY, it is reasonable to assume that this is valid for tropical forests only. For example, we expect temperate forests to be affected by more complex changes throughout the year, possibly requiring additional input information to the model, such as the DOY itself or the time of acquisition, to remain unbiased.

Similarly, the unbiased estimation with respect to the height of ambiguity and the local incidence angle is extremely relevant for large-scale applications using single-pass InSAR, as it suggests that our model is capable of delivering spatially consistent estimates, independently of the SAR and InSAR geometries.

A performance comparison of our method with respect to the state of the art in the literature is also of interest. Compared to the preliminary work published in Carcereri et al. (2023), the newly presented approach yields an overall improvement across all considered metrics. The total bias has improved by 1.36 m, the MAE by 0.30 m, and the RSME by 0.61 m. When comparing our methodology to the physical-based models, the RVoG model represents the most investigated approach (Papathanassiou and Cloude, 2001; Cloude and Papathanassiou, 2003; Guliaev et al., 2021; Chen et al., 2016; Olesk et al., 2016). Given the features used in our proposed method, we compare our performance with the inverted sinc-approximation of the RVoG model (Olesk et al., 2016), since it also only requires information about the acquisition geometry (i.e., θ_{inc} and h_{amb}) and the volumetric decorrelation coefficient γ_{vol} from a single-pol, single-baseline acquisition. The RVoG achieves an overall ME of −2.24 m, a MAE of 8.60 m and an RMSE of 10.85 m. In Guliaev et al. (2021), the proposed RVoG inversion scheme using a combination of TanDEM-X imagery and LiDAR profiles achieved an RMSE of 8.16 m and a r^2 value of 0.16 over the site of Lopé. For comparison, with our approach we achieve an RMSE of 4.88 m and r^2 of 0.50. In Denbina et al. (2018), the RVoG is inverted using multi-baseline, quad-pol acquisitions and selecting the optimal baseline using a support vector machine (SVM) trained on sparse LiDAR measurements. These experiments lead to an RMSE of 5.64 m over Pongara, of 4.99 m over Mondah, and of 5.99 m over Lopé, respectively. Using our proposed approach, we achieve an RMSE of 5.45 m, 4.00 m and 4.88 m, respectively. Finally, in Lang et al. (2019) the authors proposed a deep learning approach, which estimates the CHM values from Sentinel-2 multi-spectral data. The analyses over the AfriSAR Campaign test sites achieve a MAE of 4.9 m and an RMSE of 6.5 m when considering the yearly least cloudy acquisitions, and a MAE of 4.3 m and an RMSE of 5.6 m when applying a temporal median filter across a one year inference stack. In this context, our proposed method achieves extremely competitive results, at the advantage of requiring only a single TanDEM-X acquisition as input.

Finally, the proposed method shows a significant potential for generating multi-temporal, time-tagged products and for monitoring forest height changes in time. On the one hand, clear cuts and afforestation can be easily identified since they represent abrupt changes. On the other hand, the challenge is to monitor forests dynamics, whose variations lie within the current uncertainty boundaries of the model. To this aim, further validation is required to assess the reliability of the derived model with respect to reference data acquired at different times.

6. Conclusions

In this work, we presented a novel supervised deep learning approach for country-scale forest height estimation from single-pass TanDEM-X SAR and InSAR products. The method was trained and tested using the rasterized airborne LVIS LiDAR measurements, acquired in the context of the 2016 NASA/ESA AfriSAR campaign in Gabon. The deployment at large-scale posed a series of challenges, mainly related to missing representations of the input predictor space in the training set and to the assessment of the model reliability where no reference data is available for precise validation. To cope with these challenges, we proposed a novel model reliability measure,

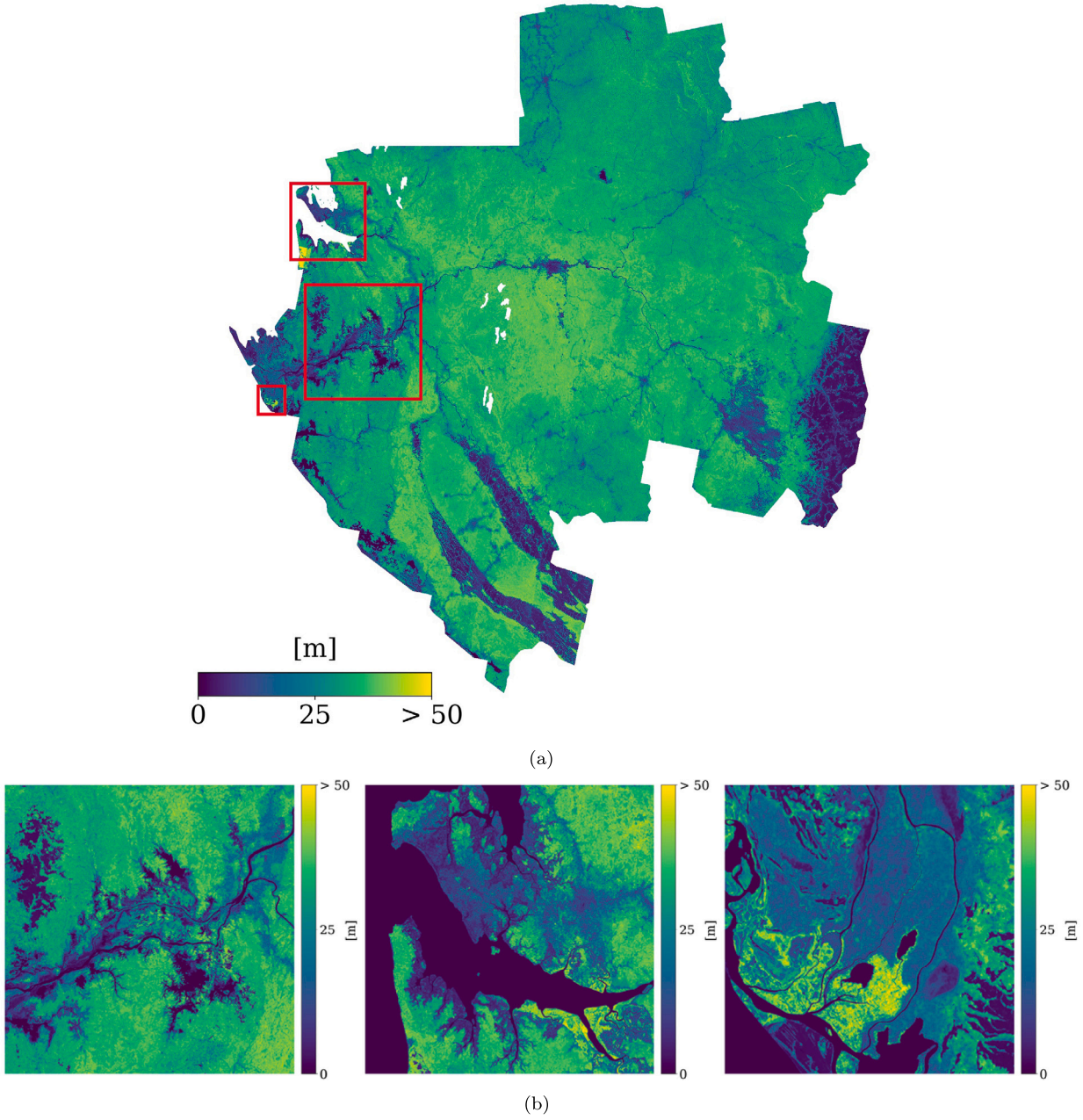


Fig. 11. (a) Country-scale mosaic of Gabon representing the CHM, generated using TanDEM-X acquisitions from the first global covered of the mission (Dec. 2010 - end of 2011). (b) Zoom-ins of the three regions included in the red boxes of the country-scale CHM mosaic. Invalid values, caused by either shadow and layover or by the unreliability of the model, are depicted in white.

called map of applicability, and we used it to drive the definition of a robust dataset for training, concentrating on the role of the height of ambiguity and of the raw DEM as input predictors. The final model delivers accurate height estimates, which show a very competitive performance with respect to state of the art methods, at the advantage of requiring only one single TanDEM-X acquisition, i.e., considering only a single baseline for each pixel. Finally, we deployed our proposed approach to map the entirety of Gabon at 25 m resolution using time-tagged data from the first global coverage of TanDEM-X acquisitions. The proposed method represents a solid starting point for setting up a reliable framework for the generation of large-scale products of biophysical forest parameters over tropical forests. As an outlook to future activities, we aim at further assessing the potential of the methodology for monitoring changes in time in the canopy height, as well as improving the model itself, by increasing its complexity to simultaneously encompass multiple forest scenarios across different continents. In order to further improve the performance, we consider to expand the framework to a multi-source approach, in which we take advantage of the synergistic use of both SAR, InSAR and multi-spectral information. Finally, we also aim at expanding our model to complementary forest parameters, such as forest coverage and above-ground biomass, thanks to the flexibility of deep learning to transfer knowledge between similar domains.

CRediT authorship contribution statement

Daniel Carcereri: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Paola Rizzoli:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Luca Dell'Amore:** Writing – review & editing, Data

curation. **José-Luis Bueso-Bello:** Writing – review & editing, Data curation. **Dino Ienco:** Writing – review & editing, Conceptualization. **Lorenzo Bruzzone:** Writing – review & editing, Supervision.

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.

Data availability

The TanDEM-X data can be accessed through the submission of a scientific proposal at: https://tandemx-science.dlr.de/. The LVIS and GEDI dataset are freely available.

Acknowledgments

The authors would like to thank the anonymous reviewers, whose valuable comments significantly helped improving the quality of the paper.

References

Becker, A., Russo, S., Puliti, S., Lang, N., Schindler, K., Wegner, J.D., 2023. Country-wide retrieval of forest structure from optical and SAR satellite imagery with deep ensembles. ISPRS J. Photogramm. Remote Sens. 195, 269–286, (en).

Blair, J.B., Rabine, D.L., Hofton, M.A., 1999. The Laser Vegetation Imaging Sensor: a medium-altitude, digitisation-only, airborne laser altimeter for mapping vegetation and topography. ISPRS J. Photogramm. Remote Sens. 54 (2–3), 115–122, (en).

Bundeswaldinventur, 2024. Surveying the forest, https://www.bundeswaldinventur.de/en/third-national-forest-inventory/surveying-the-forest.

Carcereri, D., Rizzoli, P., Ienco, D., Bruzzone, L., 2023. A deep learning framework for the estimation of forest height from bistatic TanDEM-X data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 8334–8352, (en).

Chen, H., Cloude, S.R., Goodenough, D.G., 2016. Forest canopy height estimation using Tandem-X coherence data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (7), 3177–3188, (en).

Choi, C., Cazcarra-Bes, V., Guliaev, R., Pardini, M., Papathanassiou, K.P., Qi, W., Armston, J., Dubayah, R., 2023. Large scale forest height mapping by combining TanDEM-X and GEDI data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 1–14, (en).

Cloude, S.R., Papathanassiou, K.P., 2003. Three-stage inversion process for polarimetric SAR interferometry. IEE Proc., Radar Sonar Navig. 150 (3), 125, (en).

Denbina, M., Simard, M., Hawkins, B., 2018. Forest height estimation using multibaseline PolInSAR and sparse lidar data fusion. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (10), 3415–3433, (en).

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020. The global ecosystem dynamics investigation: High-resolution laser ranging of the Earth's forests and topography. Sci. Remote Sens. 1, 100002, (en).

FAO, 2020. Global Forest Resources Assessment 2020. FAO, (en).

Fatoyinbo, T., Armston, J., Simard, M., Saatchi, S., Denbina, M., Lavalle, M., Hofton, M., Tang, H., Marselis, S., Pinto, N., Hancock, S., Hawkins, B., Duncanson, L., Blair, B., Hansen, C., Lou, Y., Dubayah, R., Hensley, S., Silva, C., Poulsen, J.R., Labrière, N., Barbier, N., Jeffery, K., Kenfack, D., Herve, M., Bissiengou, P., Alonso, A., Moussavou, G., White, L.T.J., Lewis, S., Hibbard, K., 2021. The NASA AfriSAR campaign: Airborne SAR and lidar measurements of tropical forest structure and biomass in support of current and future space missions. Remote Sens. Environ. 264, 112533, (en).

Fayad, I., Baghdadi, N., Lahssini, K, 2022. An assessment of the GEDI lasers' capabilities in detecting canopy tops and their penetration in a densely vegetated, tropical area. Remote Sens. 14 (13), 2969, (en).

Fritz, T., Breit, H., Rossi, C., Balss, U., Lachaise, M., Duque, S., 2012. Interferometric processing and products of the TanDEM-X mission. In: 2012 IEEE International Geoscience and Remote Sensing Symposium. pp. 1904–1907, (en).

Gawlikowski, J., Ebel, P., Schmitt, M., Zhu, X.X., 2022. Explaining the effects of clouds on remote sensing scene classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 9976–9986, (en).

González, C., Bachmann, M., Bueso-Bello, J.-L., Rizzoli, P., Zink, M., 2020. A fully automatic algorithm for editing the TanDEM-X global DEM. Remote Sens. 12 (23), 3961, (en).

Guliaev, R., Cazcarra-Bes, V., Pardini, M., Papathanassiou, K., 2021. Forest height estimation by means of TanDEM-x InSAR and waveform lidar data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 3084–3094, (en).

Jucker, T., John, C., Chave, J., Antin, C., Barbier, N., Bongers, F., Dalponte, M., Van Ewijk, K.Y., Forrester, D.I., Haeni, M., Higgins, S.I., Holdaway, R.J., Iida, Y., Lorimer, C., Marshall, P.L., Momo, S., Moncrieff, G.R., Ploton, P., Poorter, L., Abd Rahman, K., Schlund, M., Sonké, B., Sterck, F.J., Trugman, A.T., Usoltsev, V.A., Vanderwel, M.C., Waldner, P., Wedeux, B.M.M., Wirth, C., Wöll, H., Murray Woods, Xiang, W., Zimmermann, N.E., Coomes, D.A., 2017. Allometric equations for integrating remote sensing imagery into forest monitoring programmes. Global Change Biol. 23 (1), 177–190, (en).

Kattenborn, Teja, Schiefer, Felix, Frey, Julian, Feilhauer, Hannes, Mahecha, Miguel D., Dormann, Carsten F., 2022. Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. ISPRS Open J. Photogramm. Remote Sens. 5, 100018, (en).

Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. arXiv, arXiv:1412.6980 [cs], (en).

Krieger, G., Moreira, A., Fiedler, H., Hajnsek, I., Werner, M., Younis, M., Zink, M., 2007. TanDEM-X: A satellite formation for high-resolution SAR interferometry. IEEE Trans. Geosci. Remote Sens. 45 (11), 3317–3341, (en).

Krieger, G., Zink, M., Bachmann, M., Bräutigam, B., Schulze, D., Martone, M., Rizzoli, P., Steinbrecher, U., Walter Antony, J., De Zan, F., Hajnsek, I., Papathanassiou, K., Kugler, F., Rodriguez Cassola, M., Younis, M., Baumgartner, S., López-Dekker, P., Prats, P., Moreira, A., 2013. TanDEM-X: A radar interferometer with two formation-flying satellites. Acta Astronaut. 89, 83–98, (en).

Lahssini, K., Baghdadi, N.., Le Maire, G., Fayad, I., 2022. Influence of GEDI acquisition and processing parameters on canopy height estimates over tropical forests. Remote Sens. 14 (24), 6264, (en).

Lang, N., Jetz, W., Schindler, K., Wegner, J.D., 2023. A high-resolution canopy height model of the Earth. Nat. Ecol. Evol. (en).

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

Liu, Weitang, Wang, Xiaoyun, Owens, John D., Li, Yixuan, 2020. Energy-based out-of-distribution detection. In: Proceedings of the 34th International Conference on Neural Information Processing Systems.

Martone, M., Rizzoli, P., Krieger, G., 2016. Volume decorrelation effects in TanDEM-X interferometric SAR Data. IEEE Geosci. Remote Sens. Lett. 13 (12), 1812–1816, (en).

Meyer, Hanna, Pebesma, Edzer, 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633, (en).

Morin, D., Planells, M., Baghdadi, N., Bouvet, A., Fayad, I., Le Toan, T., Mermoz, S., Villard, L., 2022. Improving heterogeneous forest height maps by integrating GEDI-based forest height information in a multi-sensor mapping process. Remote Sens. 14, 2079, (en).

Olesk, A., Praks, J., Antropov, O., Zalite, K., Arumäe, T., Voormansik, K., 2016. Interferometric SAR coherence models for characterization of hemiboreal forests using TanDEM-X Data. Remote Sens. 8 (9), 700, (en).

Papathanassiou, K.P., Cloude, S.R., 2001. Single-baseline polarimetric SAR interferometry. IEEE Trans. Geosci. Remote Sens. 39 (11), 2352–2363, (en).

Picard, S.-A., Laurent, N., Henry, M., 2012. Manual for Building Tree Volume and Biomass Allometric Equations from Filed Measurement To Prediction. Food and Agriculture Organization of the United Nations (FA0), Rome, en, OCLC: 931325352.

Ploton, Pierre, Mortier, Frédéric, Réjou-Méchain, Maxime, Barbier, Nicolas, Picard, Nicolas, Rossi, Vivien, Dormann, Carsten, Cornu, Guillaume, Viennois, Gaëlle, Bayol, Nicolas, Lyapustin, Alexei, Gourlet-Fleury, Sylvie, Pélissier, Raphaël, 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Commun. 11 (1), 4540, (en).

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C.E., Armston, J., Dubayah, R., Blair, J.B., Hofton, M., 2021. Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sens. Environ. 253, 112165, (en).

Rizzoli, P., Dell'Amore, L., Bueso-Bello, J.-L., Gollin, N., Carcereri, D., Martone, M., 2022. On the derivation of volume decorrelation from TanDEM-X bistatic coherence. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 3504–3518, (en).

Rizzoli, P., Martone, M., Gonzalez, C., Wecklich, C., Bräutigam, B., Borla Tridon, D., Bachmann, M., Schulze, D., Fritz, T., Huber, M., Wessel, B., Krieger, G., Zink, M., Moreira, A., 2017. Generation and performance assessment of the global TanDEM-X digital elevation model. ISPRS J. Photogramm. Remote Sens. 132, 119–139.

Saatchi, S., Chave, J., Labriere, N., Barbier, N., Réjou-Méchain, M., Ferraz, A., Tao, S., 2019. Afrisar: Aboveground Biomass for Lope, Mabounie, Mondah, and Rabi Sites, Gabon. ORNL Distributed Active Archive Center, (en).

Schwartz, Martin, Ciais, Philippe, Ottlé, Catherine, Truchis, Aurelien De, Vega, Cedric, Fayad, Ibrahim, Brandt, Martin, Fensholt, Rasmus, Baghdadi, Nicolas, Morneau, François, Morin, David, Guyon, Dominique, Dayau, Sylvia, Wigneron, Jean-Pierre, 2024. High-resolution canopy height map in the Landes forest (France) based on GEDI, Sentinel-1, and Sentinel-2 data with a deep learning approach. Int. J. Appl. Earth Obs. Geoinf. 128, 103711, (en).

Sica, F., Gobbi, G., Rizzoli, P., Bruzzone, L., 2021. Phi-Net: Deep residual learning for InSAR parameters estimation. IEEE Trans. Geosci. Remote Sens. 59 (5), 3917–3941, (en).

Touzi, R., Lopes, A., Bruniquel, J., Vachon, P.W., 1999. Coherence estimation for SAR imagery. IEEE Trans. Geosci. Remote Sens. 37 (1), 135–149, (en).

Wagner, Fabien H., Roberts, Sophia, Ritz, Alison L., Carter, Griffin, Dalagnol, Ricardo, Favrichon, Samuel, Hirye, Mayumi C.M., Brandt, Martin, Ciais, Philippe, Saatchi, Sassan, 2024. Sassan sub-meter tree height mapping of California using aerial images and LiDAR-informed U-Net model. Remote Sens. Environ. 305, 114099, (en).

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N., Xu, P., Ramoino, F., Arino, O., 2022. Esa Worldcover 10 M 2021 V200. Zenodo.

Zink, M., Moreira, A., Hajnsek, I., Rizzoli, P., Bachmann, M., Kahle, R., Fritz, T., Huber, M., Krieger, G., Lachaise, M., Martone, M., Maurer, E., Wessel, B., 2021. TanDEM-X: 10 years of formation flying bistatic SAR interferometry. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 3546–3565.