

Probabilistic Biomass Estimation with Conditional Generative Adversarial Networks ^{*}

Johannes Leonhardt^{1,2}[0000–0002–4505–5086], Lukas Drees¹[0000–0003–2052–1914],
Peter Jung³[0000–0001–7679–9697], and Ribana Roscher^{1,2}[0000–0003–0094–6210]

¹ Remote Sensing Group, University of Bonn

{jleonhardt, ldrees, ribana.roscher}@uni-bonn.de

² AI4EO Future Lab, Technical University of Munich & German Aerospace Center

³ Communications and Information Theory Chair, Technical University Berlin
peter.jung@tu-berlin.de

Abstract. Biomass is an important variable for our understanding of the terrestrial carbon cycle, facilitating the need for satellite-based global and continuous monitoring. However, current machine learning methods used to map biomass can often not model the complex relationship between biomass and satellite observations or cannot account for the estimation’s uncertainty. In this work, we exploit the stochastic properties of Conditional Generative Adversarial Networks for quantifying aleatoric uncertainty. Furthermore, we use generator Snapshot Ensembles in the context of epistemic uncertainty and show that unlabeled data can easily be incorporated into the training process. The methodology is tested on a newly presented dataset for satellite-based estimation of biomass from multispectral and radar imagery, using lidar-derived maps as reference data. The experiments show that the final network ensemble captures the dataset’s probabilistic characteristics, delivering accurate estimates and well-calibrated uncertainties.

1 Introduction

An ever-growing number of satellite missions produce vast amounts of remote sensing data, providing us with unprecedented opportunities to continuously monitor processes on the Earth’s surface. Extracting quantitative geoscientific information from these data requires functional models between the observations l and geographical variables of interest x . To this end, deep learning methods based on neural networks have recently established themselves due to their demonstrated capabilities to learn complex relationships. For applications like

^{*} This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1502/1–2022 - Projektnummer: 450058266, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2070 – 390732324 and by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab ”AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” (grant number: 01DD20001).

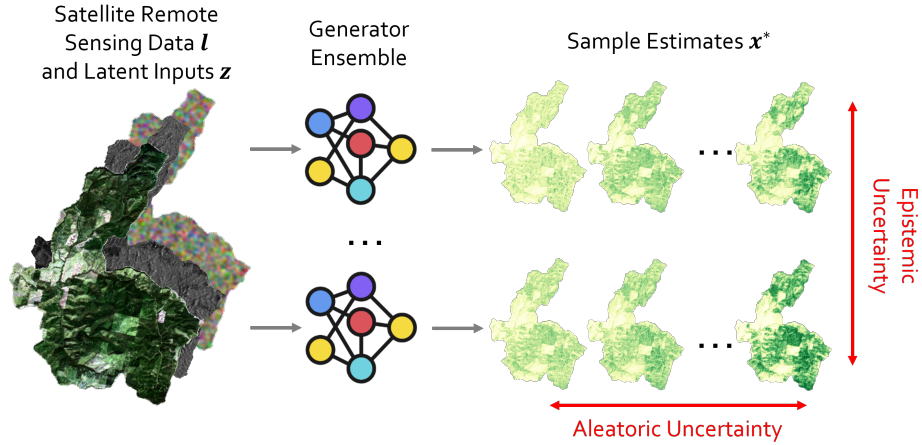


Fig. 1. Graphical summary of our methodology during inference. The observations \mathbf{l} are fed into multiple generator neural networks along with the latent codes \mathbf{z} . The sample estimates \mathbf{x}^* from each generator represent individual estimates of the predictive posterior distribution and hence, aleatoric uncertainty. The variability across the generator ensemble, on the other hand is indicative of epistemic uncertainty.

satellite-based biomass estimation, however, the inability of such methods to provide uncertainties along with their estimates represents a crucial flaw. The reason is that this problem, like many others in the field of Earth Observation, is ill-posed in the sense that there exist multiple biomass maps which are consistent with the observations. Particularly, this is due to latent variables, like tree height or tree species, which are only weakly correlated to satellite measurements but have substantial influence on biomass. This causes ambiguity, and hence, uncertainty in the estimation. While this property of our estimation task is ignored by deterministic models, probabilistic models circumvent this problem by approximating the conditional predictive posterior $P(\mathbf{x}|\mathbf{l})$ instead of point estimates. By focusing on the task of accurately approximating the predictive posterior distribution, we hope to improve the informative value of the resulting biomass products for policymaking, modeling of the carbon cycle, or other downstream applications.

In summary, we make the following contributions:

1. We motivate and describe the usage of *Conditional Generative Adversarial Networks* (CGANs) for non-parametric uncertainty quantification in biomass estimation. We point out that the variability across generated sample estimates \mathbf{x}^* is indicative of the dataset’s intrinsic *aleatoric* uncertainty, as they follow the generator’s approximation of the predictive posterior distribution.
2. We use ensembles of generator networks for capturing the *epistemic* uncertainty of CGANs, which is largely associated with instabilities of the adversarial training process. In this context, Snapshot Ensembles consisting

of generators from the same network initialization turn out to be a valid, computationally inexpensive alternative to regular ensembles.

3. We show that we can use CWGANs to easily include unlabeled data in the training process. We exploit this property to fine-tune our network to the testing data.
4. We apply an implementation of our model to a novel remote sensing dataset for satellite-based biomass estimation in Northwestern USA, evaluate it regarding the quality of the estimated predictive posterior, and show that it does not negatively affect estimation accuracy.

2 Related Work

Uncertainty Quantification in Deep Learning. Since the need for reliable and accurate uncertainty measures is not limited to problems in remote sensing, the field of probabilistic deep learning has evolved rapidly in recent years. An essential distinction in this context is between aleatoric and epistemic uncertainty: Aleatoric uncertainty is caused by the nature of the data or the underlying problem and therefore cannot be explained away, even if infinitely many training samples were available. It is therefore also an intrinsic property of ill-posed problems, where the target variable cannot be recovered from the given observations in a deterministic sense [47]. On the contrary, epistemic uncertainty is caused by limitations regarding the dataset size, the neural network’s architecture, or the optimization strategy. Therefore, this type of uncertainty can at least partly be reduced by, e.g. enlarging the dataset, specifying a more fitting architecture, or hyperparameter tuning [13,18].

For quantifying aleatoric uncertainty, multi-head neural networks, which output a parameterization of the predictive posterior – such as the mean and variance of a Gaussian distribution – have emerged as the favored technique [37]. The downside to such models is their limitation to the assumed parameterization and the resulting inability to represent the more varied predictive posteriors, which are present in real-world applications. A possible alternative is to have the neural network output distribution-free predictive intervals [39]. While this circumvents the problem of specifying a parameterization of the predictive posterior, such networks still only output individual statistics thereof so that the estimation of other moments is not possible.

One of the most popular techniques for the quantification of epistemic uncertainty, on the other hand, is ensembling where the estimation is aggregated from multiple independent neural networks [23]. Alternatives include explicitly Bayesian methods like Monte-Carlo Dropout, where dropout is applied during training and during inference [12], and Hamiltonian Monte Carlo, where parameter hypotheses are sampled by means of Markov Chain Monte Carlo and Hamiltonian dynamics [36].

Recently, multiple studies have also used CGANs and other conditional deep generative models for uncertainty quantification. Those models’ suitability for the task is motivated by their demonstrated ability to approximate highly complex (conditional) probability distributions, such as that of natural images [14,34].

In this context, the variability in the samples generated during inference is viewed as indicative of the estimation uncertainty, as they follow the approximate predictive posterior distribution. For instance, CGANs have been applied to regression and classification tasks and were observed to produce reliable uncertainties while being more stable with respect to the backbone architecture than competing methods [27]. The technique has been especially popular in the time series domain [21], where it has been used for tasks like weather fore- and nowcasting [6,40] or pedestrian [22] and aircraft [38] trajectory prediction. In traditional regression settings, CGANs have been employed, e.g. for uncertainty quantification in medical imaging [1] and atmospheric remote sensing [28].

Biomass Estimation with Remote Sensing. As one of the World Meteorological Organization (WMO)’s *Essential Climate Variables*, large-scale and continuous estimation of biomass with satellite remote sensing is important to climate scientists and policymakers [16]. Particularly, we are interested in *Aboveground Biomass* (AGB), which by definition of the United Nations Program on Reducing Emissions from Deforestation and Forest Degradation (UN-REDD), AGB denotes all “living vegetation above the soil, including stem, stump, branches, bark, seeds, and foliage” [5]. Note that hereinafter, we will use the terms biomass and AGB interchangeably. The estimation of AGB from satellite data is less cost- and labor-intensive than obtaining ground data, but poses significant challenges due to the ill-posed nature of the problem [41].

Methodologically, classical regression techniques are still commonplace in the field of biomass estimation. These range from simple linear regression [43] to geostatistical approaches [31]. Currently, random forest regression ranks among the most popular methodologies, as they turn out to be efficient, intuitive, and not significantly more inaccurate than competing methods [29,35].

Recently, however, deep learning techniques have been used with increased frequency for biomass estimation and related tasks. For instance, neural networks were shown to better estimate biomass from Landsat data than univariate regression approaches with common vegetation indices as inputs [11]. The method was subsequently investigated in light of its spatial transferability, revealing its poor generalization capabilities [10]. Significant advances in the field were the fusion of optical imagery with Synthetic Aperture Radar (SAR) data in a deep learning-based estimation approach [2] and the use of Convolutional Neural Networks (CNNs), which can better extract information from spatial patterns in the data [8]. By taking into account textural properties of the input data, CNNs have been demonstrated to especially improve estimation of vegetation properties in cases where the pixel-wise signal saturates in the presence of tall canopies [25]. Methods from probabilistic deep learning have only recently been explored for global estimation of canopy height, which is strongly correlated to biomass, using optical and spaceborne lidar data in an ensemble of multi-head networks [24,25]. In another recent work, CGANs are used to estimate spatially consistent biomass maps based on L-band SAR imagery [7]. Despite apparent similarities, this work significantly differs from ours, as the CGANs are only used deterministically and their stochastic possibilities are thus not fully exploited.

3 Methodology

We generally consider a supervised regression setup with a dataset of pairs of observations \mathbf{l} and corresponding target variables \mathbf{x} . Our approach consists of using CGANs for aleatoric uncertainty and generator Snapshot Ensembles for epistemic uncertainty. In this chapter, we will describe these two methods in detail and point out their advantages with respect to our task. Note that, for clarity, we will not explicitly distinguish between random variables and their realizations in our notation.

3.1 CGANs and Aleatoric Uncertainty

For quantifying aleatoric uncertainty, we first assume that the given samples are realizations of the conditional distribution $P(\mathbf{x}|\mathbf{l})$, which is called the predictive posterior. We furthermore assume that aleatoric uncertainty is induced by the latent variable \mathbf{z} , for which a simple prior like a standard normal distribution $P(\mathbf{z}) = \mathbb{N}_{\mathbf{0}, \mathbf{I}}$ is assumed. We may view this variable as an encoding of all factors which influence \mathbf{x} , but are inaccessible through \mathbf{l} . In the context of biomass estimation for example, \mathbf{z} encodes uncertainty about pertinent variables like tree height or density, which are only to some degree correlated to satellite measurements. Marginalization over these factors \mathbf{z} results in the model

$$P(\mathbf{x}|\mathbf{l}) = \int P(\mathbf{x}|\mathbf{l}, \mathbf{z})dP(\mathbf{z}). \quad (1)$$

Practically, this theoretical model is approximated by a CGAN: On the one hand, the generator $\mathcal{G}_\gamma(\mathbf{l}, \mathbf{z})$, parameterized as a neural network by γ , seeks to produce sample estimates of the target variable \mathbf{x} , which match the data-implied predictive posterior. On the other hand, a discriminator neural network $\mathcal{D}_\delta(\mathbf{l}, \mathbf{x})$, parameterized by δ , evaluates the generated samples by comparing them to the real samples in the training dataset and providing a suitable metric by which the generator can be optimized.

To find optimal parameter values γ^* and δ^* for the generator and the discriminator, respectively, adversarial training is employed. In the most common variants of adversarial training, the overall objective can be formalized as a mini-max game, where an objective function $L(\gamma, \delta)$ is maximized by the discriminator and minimized by the generator [14,34]:

$$\gamma^*, \delta^* = \arg \min_{\gamma} \arg \max_{\delta} L(\gamma, \delta). \quad (2)$$

Due to major practical issues with the original GAN and CGAN implementations like vanishing gradients and mode collapse [3], recent research has mostly revolved around improving the stability of adversarial training.

In particular, the Wasserstein variant of CGAN (CWGAN) aims to solve these issues by using the objective function

$$L_{\text{CWGAN}}(\gamma, \delta) = \mathbb{E}_{(\mathbf{l}, \mathbf{x})}(\mathcal{D}_\delta(\mathbf{l}, \mathbf{x})) - \mathbb{E}_{(\mathbf{l}, \mathbf{z})}(\mathcal{D}_\delta(\mathbf{l}, \mathcal{G}_\gamma(\mathbf{l}, \mathbf{z}))) \quad (3)$$

with the additional restriction $\mathcal{D}_\delta \in \mathcal{L}_1$, where \mathcal{L}_1 describes the set of Lipschitz-1 continuous functions. By virtue of the Kantorowich-Rubinstein duality, the use of this particular objective leads to the minimization of the Wasserstein-1 distance between the data-implied and the generated distribution. The favorable properties of this metric regarding its gradients with respect to γ and δ have been demonstrated to mitigate the usual issues of adversarial training when compared to the Jensen-Shannon divergence used in the original GAN [4].

In implementation, the networks are alternately optimized using stochastic gradient descent and ascent, respectively. For computing the stochastic gradients of the objective with respect to γ and δ , the expectations in equation (3) are replaced with their empirical approximation over a minibatch. At inference, we may then theoretically generate arbitrarily many sample estimates $\mathbf{x}^* = \mathcal{G}_{\gamma^*}(\mathbf{l}, \mathbf{z})$ to approximate the predictive posterior distribution by repeatedly running generator forward passes with different \mathbf{z} -inputs, sampled from the latent prior. CGANs therefore allow for the non-parametric and distribution-free modeling of aleatoric uncertainty, setting them apart from multi-head neural networks [37], where one is limited to the Gaussian parameterization. At the same time, they still approximate the full predictive posterior instead of single output statistics, as is the case for prediction intervals [39]. Instead, the sample estimates may be used to compute an approximation of a wide range of statistics of the predictive posterior. This also includes correlations in multi-output setups, which are entirely disregarded by the other methods. The accuracy of these approximations, however, may be limited by the number of the generated samples, which is subject to computation time and memory constraints.

Another advantageous aspect about using CGANs for probabilistic regression is the possibility to use unlabeled data at training time to train the generator network. This way, the model can be tuned not only with respect to the training data, but also the testing data without needing access to the corresponding labels. We expect that a model trained in this manner will be less likely to overfit, leading to greater generalization capabilities.

The root cause, why this procedure is feasible, lies within the CWGAN optimization objective, i.e., the minimization of the Wasserstein-1 distance between the generated and the real distribution as provided by the discriminator. We note, that the derivatives of the objective from equation (3) with respect to the generator’s parameters are independent of any reference data \mathbf{x} : $\nabla_\gamma L_{\text{CWGAN}} = -\mathbb{E}_{(\mathbf{l}, \mathbf{z})}(\nabla_\gamma \mathcal{D}_\delta(\mathbf{l}, \mathcal{G}_\gamma(\mathbf{l}, \mathbf{z})))$ [4, Theorem 3]. This is based on the fact, that the minimization itself takes place with respect to the joint, rather than the conditional space of \mathbf{l} and \mathbf{x} [1]. Practically, we can therefore produce sample estimates from unlabeled data and still use the discriminator – which must still be trained on the labeled training dataset – in order to evaluate them to adjust γ accordingly. We point out that this is, in fact not just a special property of CWGAN, but is true for most variants of conditional adversarial training.

3.2 Generator Ensembles and Epistemic Uncertainty

The above described model accounts for aleatoric uncertainty in regression tasks by being able to sample from its approximation of the predictive posterior distribution, but does not model the epistemic uncertainty that arises from misspecifications of the network architecture or the optimization strategy. This is overlooked by previous works who appear to assume that CGANs capture both components of uncertainty. More specifically, the likely cause of epistemic uncertainty of our generative model is the generator’s incapacity to replicate the target distribution or instabilities in the adversarial training procedure, both of which cause uncertainty in the determination of γ^* . Thus, we must not only marginalize over \mathbf{z} , which is responsible for aleatoric uncertainty, but also over different hypotheses for γ^* . In this context, the set of optimal parameters is interpreted as a random variable, as well. For simplicity, we will simply call its distribution $P(\gamma^*)$, omitting the fact that this is actually also a conditional distribution based on the above-described factors. This extends the model in equation (1) to

$$P(\mathbf{x}|\mathbf{l}) = \iint P(\mathbf{x}|\mathbf{l}, \mathbf{z}, \gamma^*)dP(\mathbf{z})dP(\gamma^*). \quad (4)$$

The resulting model effectively averages over possible optimal parameters, which results in so-called Bayesian Model Averages, of which ensembles represent one possible implementation [23,46]. In our CWGAN realization of the theoretical model, each of the generators in the ensemble thus represents one individual approximation of the predictive posterior distribution. Their combination can hence be seen as a mixture model with equal weight given to each individual generator. Statistics of the predictive posterior can be approximated by again aggregating multiple sample estimates \mathbf{x}^* , which in the combined model stem from multiple generators instead of just one.

For regular neural network ensembles, each network is initialized and trained independently from scratch. However, the training process of CGANs, and especially that of CWGANs, is time-expensive, making such a procedure impractical. We therefore turn to Snapshot Ensembles of generators, which allows for training an ensemble of networks based on a single initialization [17]. After an initial phase of T iterations of regular training with a constant learning rate λ_{max} , a cyclic learning rate schedule, particularly Cosine Annealing with Warm Restarts [30] is employed. For each cycle of T_{cyc} training iterations in this schedule, the learning rate λ_t at iteration t within the cycle is computed as

$$\lambda_t = \frac{\lambda_{max}}{2} \left(1 + \cos \left(\frac{t\pi}{T_{cyc}} \right) \right). \quad (5)$$

At the end of each cycle, the network is saved as one element of the ensemble and the learning rate is reset to λ_{max} . We believe that such an approach is especially suitable for the quantification of epistemic uncertainty of CGANs, because adversarial training is known to be unstable and oscillate around the optimum instead of converging to an equilibrium [33,45].

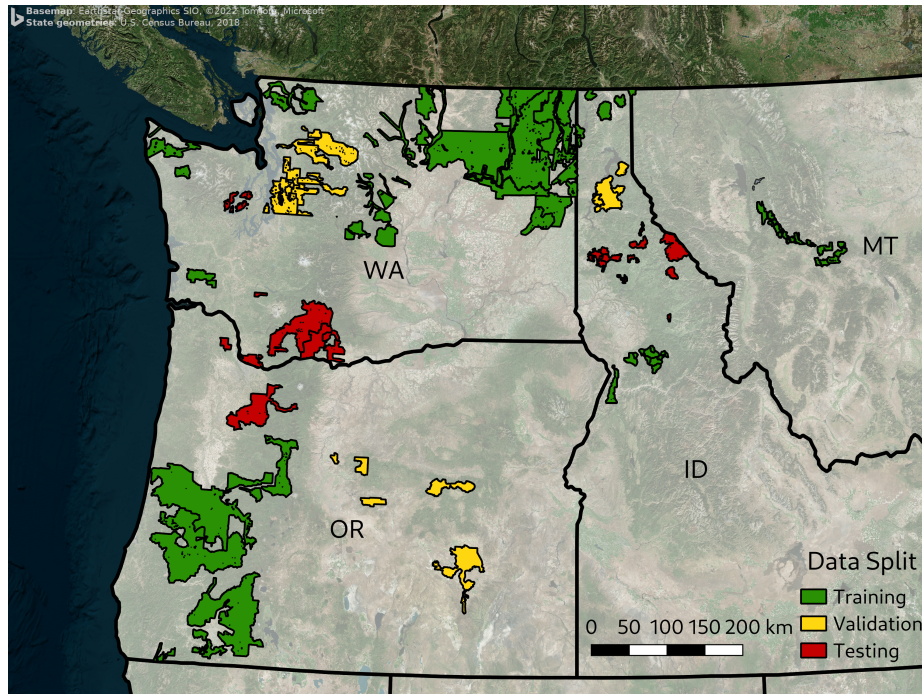


Fig. 2. Overview of the sites in our dataset and the split between training, validation and testing. For each of the sites, multispectral and SAR imagery, as well as the reference biomass map from ALS are given. The image was created using data from Bing Maps, the US Census Bureau, and the ALS reference dataset [9].

4 Application to Biomass Estimation

The next section presents the application of our methodology to the task of biomass mapping. First, we present a new dataset for satellite-based AGB estimation and afterwards apply a CWGAN implementation of our model and compare it to deterministic and multi-head neural networks in terms of the quality of the uncertainties and estimation accuracy. The code for training the models and a sample of the processed dataset are available at github.com/johannes-leonhardt/probabilistic-biomass-estimation-with-cgans-public.

4.1 Dataset

Our dataset is based on biomass maps from the US Carbon Monitoring System, which are based on airborne laserscanning (ALS) campaigns for 176 sites in Northwestern USA between 2002 and 2016, accessed through ORNL DAAC [9]. Those records are associated with multispectral imagery from the Landsat-8 satellite with its seven surface reflectance bands on the one hand and L-band

SAR imagery from ALOS PALSAR-2 with HH- and HV-polarizations and the incident angle on the other. Both satellite products were subject to several pre-processing steps like atmospheric and slope corrections and were accessed through Google Earth Engine. It has been shown, that a combination of these two sensors leads to improved estimators of vegetation characteristics, because the optical signal is sensitive towards photosynthetic parts of vegetation, while SAR backscatter values correlate with physical forest properties like tree stand height. Regarding the latter, low-frequency radars are preferred, as they are to penetrate the canopy more deeply [29]. Another advantage of data from SAR and optical sensors is their global availability over long timespans. ALS records, while much more laborious to obtain than satellite data, provide far better correlation with field measurements of AGB and their use as references in this particular setup is hence justified [48].

To ensure spatial consistency between the three data sources, the multispectral and SAR images are resampled to the grid of the ALS biomass records, which have a resolution of $30m$. For temporal consistency, we choose Landsat-8 and ALOS PALSAR-2 composites for each of the sites for the year the ALS data was recorded. While there exist ready-to-use composites for ALOS PALSAR-2 [44], Landsat-8 composites were created manually by taking the 25th percentile of all images from the leaf-on season (March to September) of that respective year with a cloud cover of less than 5%. For ALS biomass records from 2013 and 2014, we allow association of the SAR composites from 2015, as the composites are only available from that point onward. All records taken before 2013, however, were discarded from the dataset. The remaining 96 sites are manually divided into geographically separated training, validation and testing datasets, as depicted in Figure 2.

4.2 Implementation Details

As the backbone architecture for the CWGAN generator, we use a slightly modified variant of U-Net [42]. Besides the standard convolutions, we use strided convolutions in the contracting path and strided transposed convolutions in the expansive path. All hidden layers are activated with leaky ReLU, while the final output layer uses ReLU to enforce positive biomass estimates. For the Wasserstein discriminator, we use a CNN backbone with layers consisting of convolutions, strided convolutions and leaky ReLU activations. Lastly, a single, unactivated linear output layer is applied. The inputs for both networks are concatenations of the respective input tensors along the channel dimension, i.e. \mathbf{l} and the three-dimensional \mathbf{z} -inputs for the generator and \mathbf{l} with either \mathbf{x} or \mathbf{x}^* for the discriminator.

As is usual for CWGAN training, we perform five discriminator update for each generator update. For enforcing the Lipschitz-1 constraint in the discriminator, weight clipping to the range $[-0.01; 0.01]$ is used [4]. We additionally find that it is useful to pre-train the generator deterministically on MSE to find an initialization before the subsequent adversarial training. Afterwards, snapshot

Table 1. List of hyperparameters of the best performing models.

Method	T_{pt}	T	T_{cyc}	λ_{max}	Normalization
Deterministic	N/A	2000	N/A	5×10^{-5}	BN
Multi-Head	N/A	5000	1000	5×10^{-5}	BN
CWGAN	2000	8000	1000	1×10^{-5}	None

ensembles of 10 networks are then trained according to the above described procedure. For examining the advantages of using unlabeled data during training, as described above, we train another CWGAN in the same manner but sample the generator minibatches from the testing dataset, rather than the training dataset during the snapshot ensembling phase. We would like to stress again that access to the labels is not required when training the generator, and this approach is therefore implementable in practice to specifically fine-tune the generator to the dataset it shall later be applied to.

For comparison, a snapshot ensemble of multi-head neural networks is trained on an adapted variant of the MSE loss [20]. We also report results of a network that has been trained deterministically to minimize MSE as a standard regression baseline. For both the multi-head and the deterministic baselines, we use the same backbone U-Net architecture as for the CWGAN generator.

In all cases, training is performed on minibatches of 128 patches of 64×64 pixels, which are sampled from random positions in the training maps at each training iteration. Since U-Net is fully convolutional, however, the network can be applied to inputs of arbitrary size at inference. The validation dataset is used to optimize the hyperparameters individually for each method. Particularly, we conduct a search over the maximum learning rate $\lambda_{max} \in [1 \times 10^{-4}, 5 \times 10^{-5}, 1 \times 10^{-5}]$, find a suitable number of pre-training (only in the case of CWGAN), regular training and cycle iterations, T_{pt} , T and T_{cyc} , and decide whether to apply Batch Normalization (BN), Instance Normalization (IN), or neither of those in the U-Net backbone. The hyperparameters of the best performing model for each methodology are listed in Table 4.2.

4.3 Experimental Results

We finally evaluate the trained networks on the testing dataset. The input and reference data, as well as results of the multi-head approach and CWGAN sample estimates for one particular test site are depicted in Figure 3.

Both methods indicate uncertainties of up to about $100t/ha$ as measured by the standard deviations in the high biomass regime of about $> 350t/ha$. For larger values, we also observe that the estimation of the predictive posterior does not significantly change for either model. We interpret this as an indication of signal saturation as there are no detectable correlations between the satellite observations and biomass. This threshold is in line with that reported in other studies on L-band SAR for biomass estimation [19,32,41].

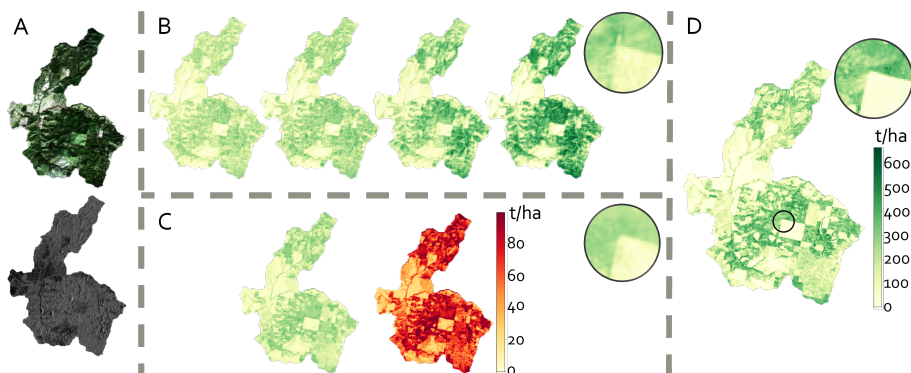


Fig. 3. Visualization of illustrative results for the test site “Big Sand Creek” in Northern Idaho. **A:** The input data, Landsat-8 in an RGB-Visualization and ALOS PALSAR-2 HH-backscatter, **B:** four sample estimates from a CWGAN, **C:** predicted Gaussian mean and standard deviation from a multi-head neural network, and **D:** the corresponding ALS-derived reference biomass map [9].

For quantitative evaluation of the estimated predictive posterior distributions, we use the Quantile Calibration Error

$$\text{QCE} = \frac{1}{M} \sum_{m=1}^M |F(q_m) - q_m|. \quad (6)$$

This metric is derived from calibration plots [20], which describe the frequencies $F(q)$ of reference values lying within quantile q of the predicted distribution. In this context, M refers to the number of regularly spaced quantile values q_m , for which the frequency is evaluated. To determine the quantiles, the estimated cumulative distributions are evaluated at the reference values. In the case of a well-calibrated predictive posterior, the reference value should be equally likely to fall into each quantile of the predicted distribution, such that the calibration line is close to the diagonal $F(q) = q$. Intuitively, QCE describes the approximate area between this ideal diagonal and the actual calibration line. This way of quantifying calibration is preferred over other common metrics like the Expected Calibration Error [15], which only take into account a single uncertainty statistic. In contrast, QCE allows for evaluating the full approximated predictive distribution including its overall shape, making it sensitive towards possible misspecifications of the uncertainty’s parametric model.

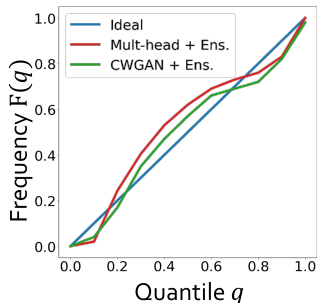
Additionally, the Root Mean Squared Error

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^N \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|^2} \quad (7)$$

measures the accuracy across N point estimates $\hat{\mathbf{x}}_n$, as usual in a deterministic regression task. While the point estimate is directly provided in the multi-head

Table 2. Quantitative results for QCE and RMSE of the different methods on the testing dataset. Ensembling (Ens.) refers to the use of Snapshot Ensembles as described above. Fine-tuning (F.-T.) denotes our method for training time usage of unlabeled test data data. Exemplary calibration plots used to calculate QCE are shown in the adjacent figure. For non-ensembles, the given values are averages over the metrics of the individual networks in the ensemble.

Method	QCE [-]	RMSE [t/ha]
Deterministic	N/A	86.15
Multi-Head	0.0853	85.12
+ Ensembling	0.0779	83.23
CWGAN	0.0851	87.67
+ Ensembling	0.0657	85.38
+ Fine-tuning	0.0889	89.92
+ Ens. + F.-T.	0.0610	86.19



setting as the distribution mean, it is computed as the average over the generator’s sample estimates \mathbf{x}^* in the case of CWGAN. Note that by using probabilistic methods we do not primarily seek to improve RMSE, but only use it to verify that the usage of a probabilistic framework does not negatively affect estimation accuracies.

Results are reported in Table 2. For CWGANs, the values are computed based on 50 sample estimates \mathbf{x}^* from each generator. We view this number as sufficient as the metrics differ only insignificantly across multiple evaluation runs. As our main result, we observe that ensembles of CWGANs produce slightly better calibrated uncertainties than those of multi-head networks. The Snapshot Ensembling methodology is able to improve the calibration in both methods. However, while this improvement is marginal in multi-head networks, it is more significant in CWGANs and individual CWGAN generators do indeed not provide better estimates of the predictive posterior than individual multi-head networks. This insight is consistent with our expectation that Snapshot Ensembles are especially helpful in the context of CGAN-based methodologies, because fluctuations within the adversarial training process are successfully averaged out. For the multi-head approach on the other hand, snapshot ensembling cannot overcome the misspecification of the predictive posterior’s parameterization.

For CWGANs, calibration is slightly improved when using our fine-tuning methodology, supporting our claim that such procedures may be helpful for training models which are tailored for application to specific data, e.g., from a particular geographical region. This approach does, however, come with the additional computational cost of retraining the network every time it is applied to a new dataset and more research may be needed to determine if the method is able to consistently improve CWGAN-based estimation.

Furthermore, it was demonstrated that both probabilistic approaches do not suffer from a significant loss in accuracy when compared to deterministic methods. In fact, our results show that snapshot ensembling consistently reduces RMSE by averaging over individual networks' epistemic uncertainties. On a final note, we observe that CWGANs deliver more consistent biomass maps than point estimators by being sensitive towards correlations in the output. This can be observed from the small scale details of the maps in Figure 3: Whereas the point estimates of multi-head networks are rather smooth, the texture of the reference maps is at least partly replicated in the CWGAN sample estimates.

5 Conclusion and Outlook

This paper presented a new approach to uncertainty quantification in satellite-based biomass estimation. In particular, we used CGANs for non-parametric approximation of aleatoric uncertainty and Snapshot Ensembles for quantifying epistemic uncertainty. The methods were discussed theoretically, implemented and evaluated on a novel dataset consisting of optical and SAR imagery and ALS-derived references. The experiments demonstrated that our method is competitive with the commonly used parametric multi-head approach without loss in accuracy.

In light of these promising results, we envision several future research directions. From a methodological standpoint, we hope to encourage future works at the intersection of deep generative models and uncertainty quantification. Beyond biomass estimation, we consider investigations of CGANs' capabilities for uncertainty quantification in other remote sensing regression problems with similar restrictions, or even tasks from different domains like classification and segmentation to be interesting topics of future studies. Moreover, we believe that our approach to training time usage of unlabeled data is worthy of more detailed and fundamental investigation as further analyses may pave the way for general applications in the context of semi-supervised learning or domain adaptation.

Our dataset offers a starting point for the inclusion of data from more satellite missions and globally distributed biomass reference records. Such a dataset in combination with multi-sensor, probabilistic estimation methods like ours would enable the creation of reliable global and multitemporal biomass monitoring products. To improve the overall accuracy of such products, we also look forward to new spaceborne sensor technologies, such as the P-band SAR onboard the BIOMASS mission [26], which is set to launch in 2023.

References

1. Adler, J., Öktem, O.: Deep Bayesian Inversion (2018), arXiv.org e-Print 1811.05910
2. Amini, J., Sumantyo, J.T.S.: Employing a Method on SAR and Optical Images for Forest Biomass Estimation. *IEEE Transactions on Geoscience and Remote Sensing* **47**(12), 4020–4026 (2009)

3. Arjovsky, M., Bottou, L.: Towards Principled Methods for Training Generative Adversarial Networks. In: International Conference on Learning Representations (2016)
4. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein Generative Adversarial Networks. In: International Conference on Machine Learning. pp. 214–223 (2017)
5. Ashton, M.S., Tyrrell, M.L., Spalding, D., Gentry, B.: Managing Forest Carbon in a Changing Climate. Springer Science & Business Media (2012)
6. Bihlo, A.: A generative adversarial network approach to (ensemble) weather prediction. *Neural Networks* **139**, 1–16 (2021)
7. Björk, S., Anfinsen, S.N., Næsset, E., Gobakken, T., Zahabu, E.: Generation of Lidar-Predicted Forest Biomass Maps from Radar Backscatter with Conditional Generative Adversarial Networks. In: International Geoscience and Remote Sensing Symposium. pp. 4327–4330 (2020)
8. Dong, L., Du, H., Han, N., Li, X., Zhu, D., Mao, F., Zhang, M., Zheng, J., Liu, H., Huang, Z., He, S.: Application of Convolutional Neural Network on Lei Bamboo Above-Ground-Biomass (AGB) Estimation Using Worldview-2. *Remote Sensing* **12**(6), 958 (2020)
9. Fekety, P.A., Hudak, A.T.: LiDAR Derived Forest Aboveground Biomass Maps, Northwestern USA, 2002-2016. Oak Ridge National Laboratory Distributed Active Archive Center (2020)
10. Foody, G.M., Boyd, D.S., Cutler, M.E.J.: Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. *Remote Sensing of Environment* **85**(4), 463–474 (2003)
11. Foody, G.M., Cutler, M.E., McMorrow, J., Pelz, D., Tangki, H., Boyd, D.S., Douglas, I.: Mapping the biomass of Bornean tropical rain forest from remotely sensed data. *Global Ecology and Biogeography* **10**(4), 379–387 (2001)
12. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: International Conference on Machine Learning. pp. 1050–1059 (2016)
13. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X.: A Survey of Uncertainty in Deep Neural Networks (2021), arXiv.org e-Print 2107.03342
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Conference on Neural Information Processing Systems. pp. 2672–2680 (2014)
15. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On Calibration of Modern Neural Networks. In: International Conference on Machine Learning. pp. 1321–1330 (Jul 2017)
16. Houghton, R.A., Hall, F., Goetz, S.J.: Importance of biomass in the global carbon cycle. *Journal of Geophysical Research* **114**, G00E03 (2009)
17. Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q.: Snapshot Ensembles: Train 1, Get M for Free. In: International Conference on Learning Representations (2017)
18. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning* **110**(3), 457–506 (2021)
19. Joshi, N., Mitchard, E.T.A., Brolly, M., Schumacher, J., Fernández-Landa, A., Johannsen, V.K., Marchamalo, M., Fensholt, R.: Understanding ‘saturation’ of radar signals over forests. *Scientific Reports* **7**(1), 3505 (2017)

20. Kendall, A., Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: Conference on Neural Information Processing Systems. pp. 5580–5590 (2017)
21. Koochali, A., Schichtel, P., Dengel, A., Ahmed, S.: Probabilistic Forecasting of Sensory Data With Generative Adversarial Networks – ForGAN. *IEEE Access* **7**, 63868–63880 (2019)
22. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, S.H., Savarese, S.: Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks. In: Conference on Neural Information Processing Systems. pp. 137–146 (2019)
23. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In: Conference on Neural Information Processing Systems. pp. 6405–6416 (2017)
24. Lang, N., Jetz, W., Schindler, K., Wegner, J.D.: A high-resolution canopy height model of the Earth (2022), arXiv.org e-Print 2204.08322
25. Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., Wegner, J.D.: Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. *Remote Sensing of Environment* **268**, 112760 (2022)
26. Le Toan, T., Quegan, S., Davidson, M.W.J., Balzter, H., Paillou, P., Papathanassiou, K., Plummer, S., Rocca, F., Saatchi, S., Shugart, H., Ulander, L.: The BIOMASS mission: Mapping global forest biomass to better understand the terrestrial carbon cycle. *Remote Sensing of Environment* **115**(11), 2850–2860 (2011)
27. Lee, M., Seok, J.: Estimation with Uncertainty via Conditional Generative Adversarial Networks. *Sensors* **21**(18), 6194 (2021)
28. Leinonen, J., Guillaume, A., Yuan, T.: Reconstruction of Cloud Vertical Structure With a Generative Adversarial Network. *Geophysical Research Letters* **46**(12), 7035–7044 (2019)
29. Li, Y., Li, M., Li, C., Liu, Z.: Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. *Scientific Reports* **10**, 9952 (2020)
30. Loshchilov, I., Hutter, F.: SGDR: Stochastic Gradient Descent with Warm Restarts. In: International Conference on Learning Representations (2016)
31. Maselli, F., Chiesi, M.: Evaluation of statistical methods to estimate forest volume in a mediterranean region. *IEEE Transactions on Geoscience and Remote Sensing* **44**(8), 2239–2250 (2006)
32. Mermoz, S., Réjou-Méchain, M., Villard, L., Le Toan, T., Rossi, V., Gourlet-Fleury, S.: Decrease of L-band SAR backscatter with biomass of dense forests. *Remote Sensing of Environment* **159**, 307–317 (2015)
33. Mescheder, L., Geiger, A., Nowozin, S.: Which Training Methods for GANs do actually Converge? In: International Conference on Machine Learning. pp. 3481–3490 (2018)
34. Mirza, M., Osindero, S.: Conditional Generative Adversarial Nets (2014), arXiv.org e-Print 1411.1784
35. Mutanga, O., Adam, E., Cho, M.A.: High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation* **18**, 399–406 (2012)
36. Neal, R.M.: Bayesian Learning for Neural Networks. Ph.D. thesis, University of Toronto (1995)

37. Nix, D., Weigend, A.: Estimating the mean and variance of the target probability distribution. In: International Conference on Neural Networks. pp. 55–60 (1994)
38. Pang, Y., Liu, Y.: Conditional Generative Adversarial Networks (CGAN) for Aircraft Trajectory Prediction considering weather effects. In: AIAA Scitech Forum (2020)
39. Pearce, T., Brintrup, A., Zaki, M., Neely, A.: High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach. In: International Conference on Machine Learning. pp. 4075–4084 (2018)
40. Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., Mohamed, S.: Skilful precipitation nowcasting using deep generative models of radar. *Nature* **597**(7878), 672–677 (2021)
41. Rodríguez-Veiga, P., Wheeler, J., Louis, V., Tansey, K., Balzter, H.: Quantifying Forest Biomass Carbon Stocks From Space. *Current Forestry Reports* **3**, 1–18 (2017)
42. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241 (2015)
43. Roy, P.S., Ravan, S.A.: Biomass estimation using satellite remote sensing data—An investigation on possible approaches for natural forest. *Journal of Biosciences* **21**(4), 535–561 (1996)
44. Shimada, M., Ohtaki, T.: Generating Large-Scale High-Quality SAR Mosaic Datasets: Application to PALSAR Data for Global Monitoring. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **3**(4), 637–656 (2010)
45. Wang, Y., Zhang, L., van de Weijer, J.: Ensembles of Generative Adversarial Networks. In: Conference on Neural Information Processing Systems (2016), workshop on Adversarial Training
46. Wilson, A.G., Izmailov, P.: Bayesian Deep Learning and a Probabilistic Perspective of Generalization. In: Conference on Neural Information Processing Systems. pp. 4697–4708 (2020)
47. Zhang, C., Jin, B.: Probabilistic Residual Learning for Aleatoric Uncertainty in Image Restoration (2019), arXiv.org e-Print 1908.01010v1
48. Zolkos, S.G., Goetz, S.J., Dubayah, R.: A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. *Remote Sensing of Environment* **128**, 289–298 (2013)