EAEEE 4000 – Final Project

Creating Global Coverage for Biomass Estimation using Sentinel 2 Imagery and GEDI LIDAR Biomass Observations

Oliver Hegi (omh2115)

Fall Semester 2022

Abstract

Biomass estimations are an important tool for climate policy enforcement and large-scale carbon accounting. This paper aims to create a model that can accurately predict above-ground biomass for any location on the globe using optical imagery. For this task, a convolutional neural network (CNN) model is trained with Sentinel-2 satellite imagery and GEDI LIDAR above-ground biomass observations. The Sentinel-2 imagery has a 10 by 10-meter resolution and is fed into the model to predict biomass observations, which are rasterized and have a 1km by 1km resolution. The CNN model presented in this paper predicts biomass values relatively well but tends to overestimate biomass for areas with low actual values. This could be driven by bias in the data. An extreme gradient boosting (XGB) model is used as a benchmark. The XGB model handles the data's bias better but performs worse when then the model is run on previously withheld regions. This points towards an inability of the XGB model to extrapolate beyond areas it was trained on, which makes it unsuitable for creating global coverage. The author of this paper predicts that the CNN model's performance would continue to improve with a larger data sample.[1]

[1] The code for this project can be found under the following link:
https://github.com/oliverhegi/Global_Biomass_Estimation_Using_S2_and_GEDI

1 Introduction

Climate change is driven by the earth's carbon balance (the carbon that sits in the atmosphere and contributes to the greenhouse effect versus the carbon that is sequestered in bodies of land and oceans). The IPCC catalogues five terrestrial carbon pools that involve biomass.[2] Of those, above-ground biomass (hereinafter simply referred to as biomass) is a major storage of CO_2 as it constitutes around 30% of all terrestrial carbon pools and is heavily subjected to human interference. Land-use changes such as deforestation, agriculture, and development can all dramatically change the amount of biomass that is sequestered by trees, vegetation, and peat and swamp forests. The IPCC's Sixth Assessment Report (AR6) states that the CO_2 taken up by land carbon sinks will play an increasingly important role as the atmospheric concentration of CO_2 increases.[3] As such, being able to measure biomass is an important tool for large-scale carbon accounting. Additionally, positive climate system feedback loops, such as mass forest dieback and permafrost thawing, have the potential to dramatically impact climate change and thus need to be kept track of. Biomass estimation is also a valuable tool for environmental policy. National governments pledge to emission reductions which could be fully undercut by dramatic changes in biomass. It is often up to non-governmental organizations (NGOs) to keep track of changes in biomass which has traditionally involved taking on the ground measurements.[4] These measurements are often spatially limited and can be quite costly to conduct. For example, one method is destructive biomass estimation, where a sample of trees are cut down and weighed for their biomass.[5]

Globally sequestered biomass is constantly in flux. The comparison of biomass stocks and developments between countries requires a uniform measure. It is important to be able to apply the same methodologies to compare biomass estimates between geographies and track their temporal development. One way to achieve this is to use remote sensing technology. Remote sensing uses satellite or aircraft-based sensor technologies to detect objects on the Earth's surface, which include radar, optical imagery, and LIDAR.[6] LIDAR technology seems particularly well suited for biomass estimations since it has the "ability to sample the vertical distribution of canopy and ground surfaces, providing detailed structural information about vegetation."[7] Optical imagery, is another promising technology for biomass sensing since it is more widely available and can deliver coverage over large areas at a far lower cost.

This paper uses a convolutional neural network (CNN) to extract features from satellite data in order to estimate sequestered above-ground biomass anywhere on the planet. The data is trained using a spatially limited biomass model based on GEDI LIDAR observations. A similar approach has been taken for canopy height estimation in a preprint paper by Lang, Jetz, Schindler, and Wegner.[8] They use GEDI tree height estimates to train Sentinel-2 imagery in order to create a global canopy height model.

[2] Eggleston et al., "2006 IPCC Guidelines for National Greenhouse Gas Inventories."
[3] IPCC, "Summary for Policymakers."
[4] Fraser, "Study Brings Satellite Data down to Earth for Biomass Tracking."
[5] Kumar and Mutanga, "Remote Sensing of Above-Ground Biomass."
[6] Kumar and Mutanga.
[7] Kumar and Mutanga.
[8] Lang et al., "A High-Resolution Canopy Height Model of the Earth."

This paper will first describe the methodology in detail by examining the data that is used for the study and describing the data collection process. Next the model describes the CNN model architecture and lays out the choice of hyperparameters. This is followed by a results section where the learning curve and confusion plot for the CNN are analyzed and compared to the results from an extreme gradient boosting (XGB) model. The results are then critically evaluated and further steps are considered.

2 Methodology

2.1 Data

This paper uses two types of data for the purpose of creating a model that can generate accurate above-ground estimates for any location on earth. The predictors (input data) are optical images from the Copernicus Sentinel-2 (S2) Mission from the European Space Agency. S2 has global coverage and comprises of two polar-orbiting satellites that capture a 290 km wide band and has a revisit time of 5 days (depending on latitude and cloud coverage).[9] S2 (L2A) has 13 spectral bands from the visible to the shortwave infrared.[10] This project uses the three bands for red, green, and blue (B02, B03, B04) which each have a resolution of 10 by 10 meters. Combined these three bands provide the highest resolution color representation of the surface of the earth.

The target data for the analysis are 1km by 1 km estimates of above-ground biomass density based on Global Ecosystem Dynamics Investigation (GEDI) observations.[11] GEDI is an instrument attached to the International Space Station which produces high-resolution light detection and ranging observations (LIDAR) that capture the 3-dimensional structure of the earth. GEDI is the highest quality LIDAR sensor in orbit to date and is capable of canopy-penetrating measurements to provide high quality biomass information. The dataset used for the analysis is called GEDI L4B Gridded Above Ground Biomass Density (hereinafter simply referred to as GEDI) and is based on eight concurrent tracks that consist of roughly 25 m footprint samples. The data was collected between 18.04.2019 and 04.08.2021.[12] The biomass estimate does not directly correspond to an observation but are the result of a hybrid estimation technique developed by the GEDI Science Team.[13] The researchers use an inventory of above-ground biomass density plots based on field observations and coincident airborne laser scanning.[14] Based on the hybrid estimation technique, a gridded dataset was created with 1km by 1km resolution that contains estimates for all the samples taken by the GEDI instrument.

[9] "Sentinel-2 - Missions - Sentinel Online - Sentinel Online."
[10] Sinergise, "Sentinel-2 L2A."
[11] Dubayah et al., "GEDI L4B Gridded Aboveground Biomass Density, Version 2."
[12] Dubayah et al.
[13] Land Product Validation Subgroup (Working Group on Calibration and Validation, Committee on Earth Observation Satellites), "Aboveground Woody Biomass Product Validation Good Practices Protocol."
[14] Dubayah et al., "GEDI L4B Gridded Aboveground Biomass Density, Algorithm Theoretical Basis Document."

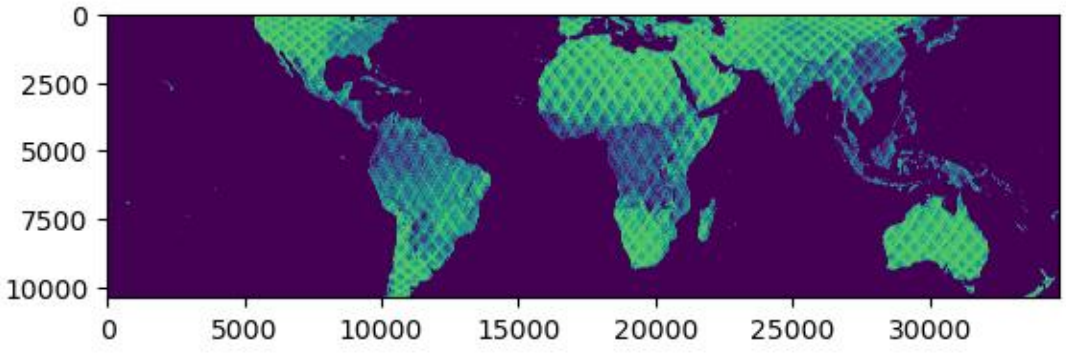

Figure 1: Visual representation of the GEDI L4B GeoTIFF (purple represents no observation and green shading is the amount of estimated biomass, with pixel count on the axes)[15]

The GEDI L4B data is contained in a 1km resolution gridded GeoTIFF (538 MB) that contains a satellite mosaic in form of a world map. As is visible in figure 1, the observations correspond to the satellite tracks covered by the GEDI instrument. This limitation results in a large degree of missingness. There is also a complete lack of data beyond the coordinates 173W, 52S, 173E, and 52N.

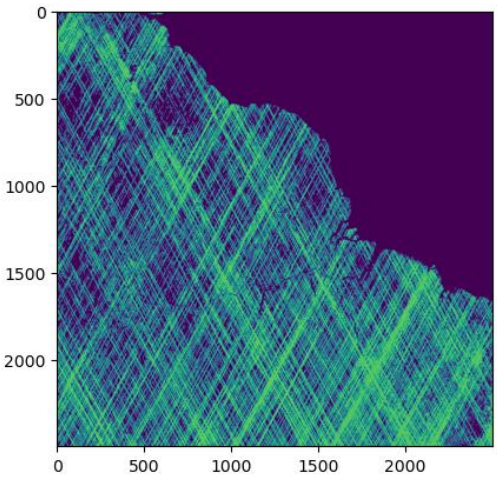

Figure 2: Sample GEDI biomass observation at 2,500km scale

Figure 2 shows a sample of the data at a 2,500 by 2,500km scale. The figure shows the green satellite tracks very clearly. The systematic missingness in the data presents a problem for the modelling approach taken in this paper. The GEDI data cannot be processed with the missing data in place. This presents two options in order to correct for the missingness. One approach could be to impute the data using an algorithm such as k-nearest neighbors (KNN). This would successfully remove the artifacts created by the satellite tracks; however, this might lead to significant bias in the data. The goal the model is to recognize differences in biomass such as where a forest ends or landscapes transition into less vegetation dense landscapes. Given the very large degree of missingness, it is unlikely the KNN would be able preserve that information in the data. The second option, and the one chosen for this study, is to process only the areas that there is data for. This could be achieved by masking the satellite tracks and removing all the missing data or by processing the data at a scale at which the missingness can be removed by selecting only samples that do not contain any missingness.

[15] Image created with Python GDAL using data from Dubayah et al., "GEDI L4B Gridded Aboveground Biomass Density, Version 2."

The approach taken in this paper is to process the GEDI data one pixel at a time. As such, a 1 km by 1 km image from S2 (with 10m resolution) is fed into the algorithm and matched with a GEDI value which represents one pixel in the GeoTIFF.

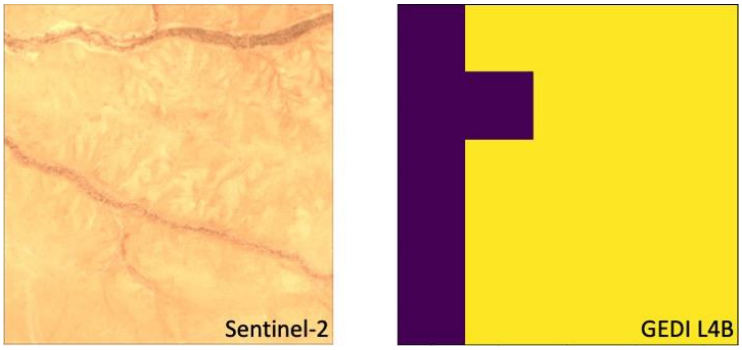

Figure 3: Sentinel-2 image with coresponding GEDI data side-by-side

Figure 3 illustrates the problem with the underlying data and the approach taken to solve it. The GEDI data in Figure 3 has a line of missing datapoints resulting from the satellite tracks. This line does not correspond to anything in the S2 image and would be problematic to include in the training data. This problem is circumvented by processing the GEDI data pixel-by-pixel and removing any pixels that do not contain any data. The corresponding S2 image will still be 1 km by 1 km and should thus contain enough information to be linked to the GEDI value.

2.2 Data Collection

The dataset used for the models contains 10,000 observations (10,000 S2 images and corresponding GEDI values) and was compiled by linking an image from S2 to a GEDI biomass estimate using coordinates. The GEDI data is stored in a GeoTIFF that stores spatial referencing information in the form of coordinates. The S2 images can be pulled via a Python API which is provided by SentinelHub.[16]

Since the degree of missingness for the GEDI GeoTIFF is high, the GeoTIFF was first sampled for a pixel by generating a random coordinate. If the pixel contains data those coordinates are then passed to the SentinelHub API to pull an image for the corresponding location. The time interval for the S2 images was restricted to the period between 18.04.2019 and 04.08.2021 which corresponds to the date range within which the GEDI data was collected. Since the date range is very broad, the Sentinel API was configured to be maximally averse to cloud cover such that the images are as unobstructed as possible.

2.3 Model

The problem being addressed by this paper is fundamentally an image processing (regression) problem. CNNs are a popular method for image processing. CNNs are artificial neural networks use filters, in place of neurons, to process an image. The resolution of the filter is typically less than the size of the image, which results in a process called convolution. The convolution layer,

[16] "Sentinel Hub."

which extracts features from the image, is then followed by an activation layer. After that, the input is pooled and ultimately feeds into a fully connected layer.[17]

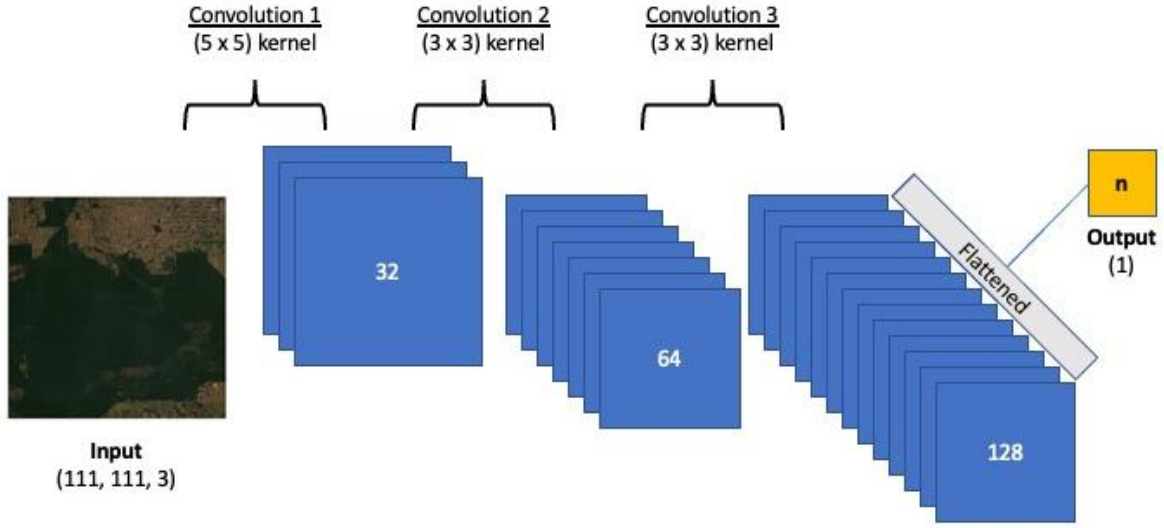

Figure 4: Illustration of the CNN configuration used[18]

As a base model for this project, a CNN architecture, as shown in Figure 4, is used. The data is fed into the model as images of dimension 111 by 111 with three channels for RGB. The model is then convoluted into 32 kernels, which have the dimension 5 by 5 and are fed into another convolution layer with 64 (3 by 3) kernels. The last layer has 128 (3 by 3) kernels. All three convolution layers use a ReLu activation function, except for the dense layer at the end which has a linear function and produces a single number which is the estimate for the GEDI biomass value. The learning rate is set at 0.1 and the batch size is 512 with a validation loss early stopping mechanism implemented at a patience of 5 epochs. This architecture is designed to convolute the 111 by 111 image into smaller filters which can pick up features from the image.

This configuration is used as baseline and is compared against a series of other hyperparameter configurations for a CNN. Other versions vary the depth of the hidden layers, the size of the filters, the activation function, the number of layers, the learning rate, and the batch size. As a benchmark an XGB model is also used.

[17] Gentine, "EAEEE4000 Machine Learning for Environmental Engineering and Science Slides."
[18] This model can be found in the Github Repository in the file CNN_V1.ipynb.

3 Results

3.1 Baseline model

The model laid out in the previous section produces the following results:

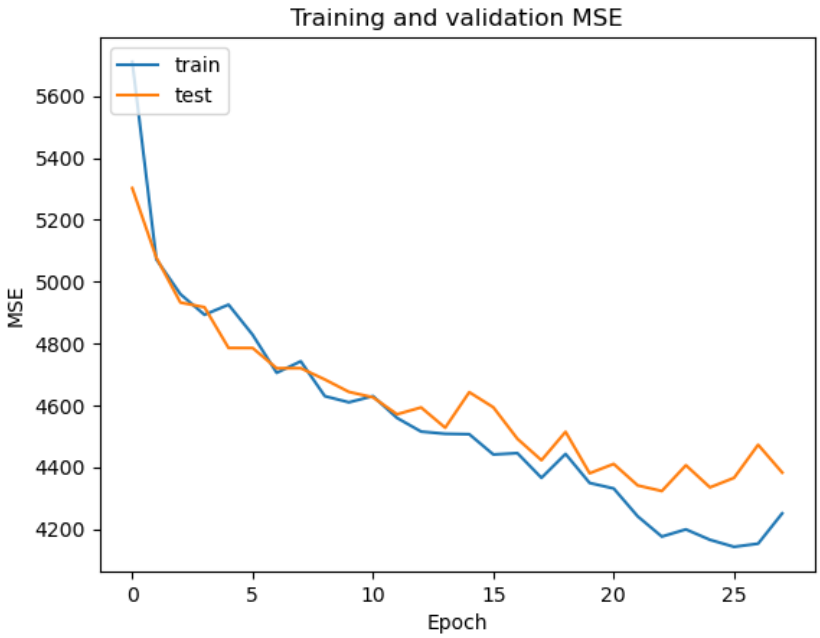

Figure 5: Learning curves for CNN (V1)

Figure 5 shows a steady reduction in mean squared error (MSE) and a relatively fast convergence towards a solution. However, at some point the model starts to overfit and is stopped by the validation loss early stopping mechanism. The MSE on test data (out-of-sample) is 4181 which is slightly below the variance of the target (5101).

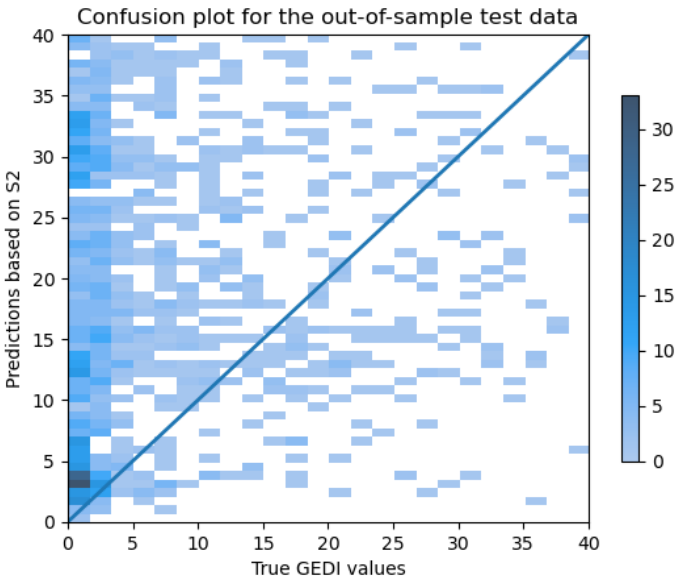

Figure 6: Confusion plot for CNN (V1)

The confusion plot displays the true test values against the out-of-sample predictions of the model. Ideally the datapoints would be clustered along the 45-degree line. Figure 6 shows that CNN (V1) performs decently well but has some systematic bias. The model consistently overestimates areas with very low biomass.

3.2 Hyperparameters

A series of other hyperparameters were tested, with differing success. The two most promising changes to the base model were reducing the batch size to 256 (CNN_V2) and adding one additional layer with 256 (1 by 1) kernels (CNN_V3). Going beyond that complexity or changing the activation functions led to unilaterally worse performance. Ultimately CNN_V1 was chosen because it most consistently produced the best balance between a low MSE and a relatively even distribution of the errors.

3.3 Extreme Gradient Boosting

XGB is a regression tree models and fits many subsequent small trees on the residuals of the previous models in order to learn the intricacies of the data. Since regression tree-based models use a series of cutoffs to separate the data, their ability to extrapolate is limited. This makes XGB less suited for this task as the main goal is to use the model to extrapolate into uncovered areas and regions that are completely outside the bounds of the training data. Nevertheless, an XGB model was performed as a benchmark.

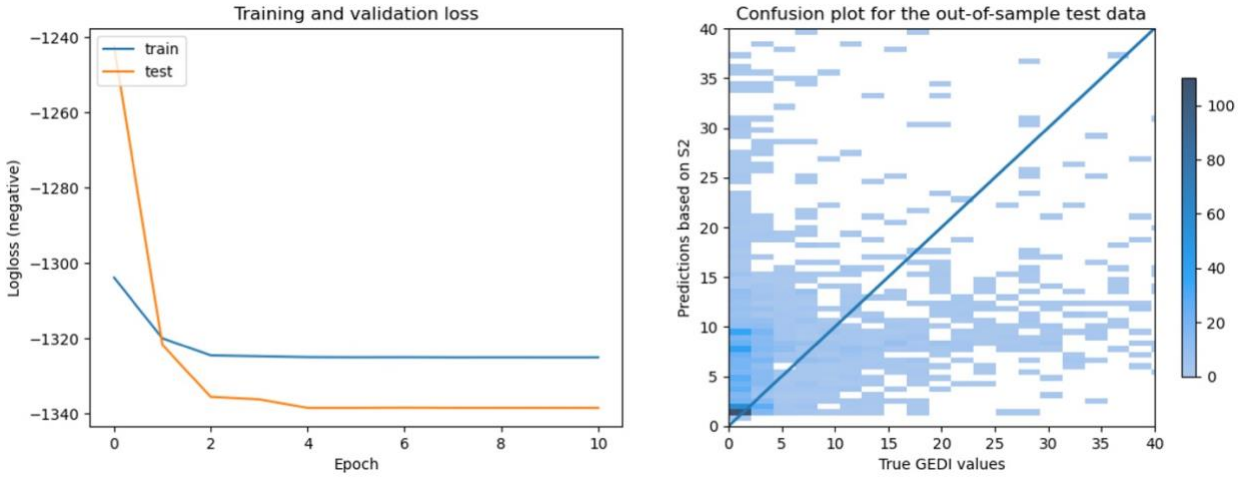

Figure 7: Learning curves and confusion plot for XGB (V1)

The learning curves and the confusion plot show that the XGB model performs very well. The predictions are far more evenly spread between both sides of the 45-degree line. In addition, the MSE is also lower at 3729. This would suggest that XGB might be a more suitable model for this problem; however, the problem with extrapolation could cause some issues.

XGB also allows one to pass a set of weights to the loss function. In this case, that ability can be used to weigh the loss more where the true value is large and deprioritize the cases where the actual biomass estimation is very low.

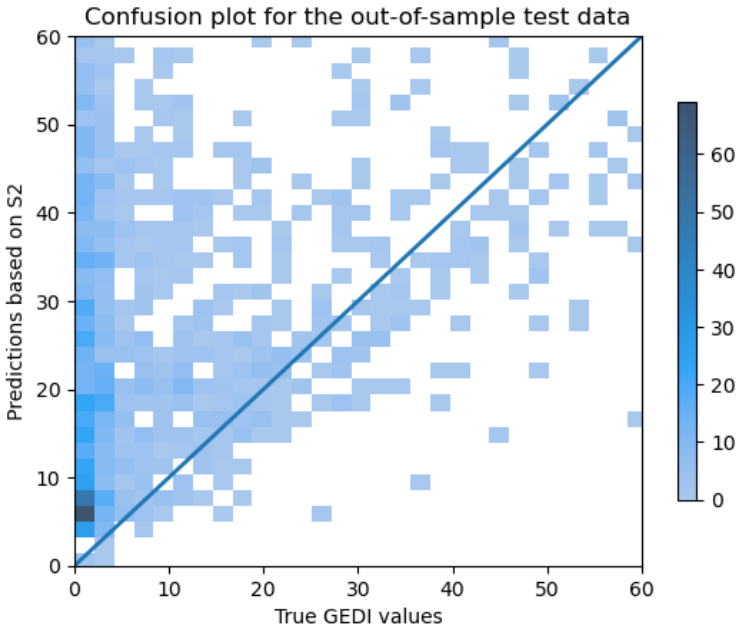

Figure 8: Confusion plot for XGB (V3)

Figure 8 shows the confusion plot for an XGB model where the loss was weighted by the corresponding y value divided by 10. This way large biomass observations would factor into the loss function by a large degree and observations with very low biomass would only factor in very slightly. The results show that are more predictions along the 45-degree line but the predictions are now even less evenly distributed.

3.4 Ability to extrapolate

The GEDI biomass coverage ends as 52 North and 52 South. As such, to create global coverage, the model will be run on latitudes it has never before been exposed to. In order to simulate how this extrapolation might play out, the most successful CNN and XGB models were both trained on a separate dataset where the training and testing data are systematically different. The training data was sampled only east of 40W longitude, and the testing data was drawn only west of 40W. This way, the entire North and South American continents are excluded from the training data and make up the entirety of the testing data. If it can be assumed that biomass accumulation in

North and South America have some systematic differences relative to the rest of the world, this data configuration will provide some additional insight into the models' abilities to extrapolate.

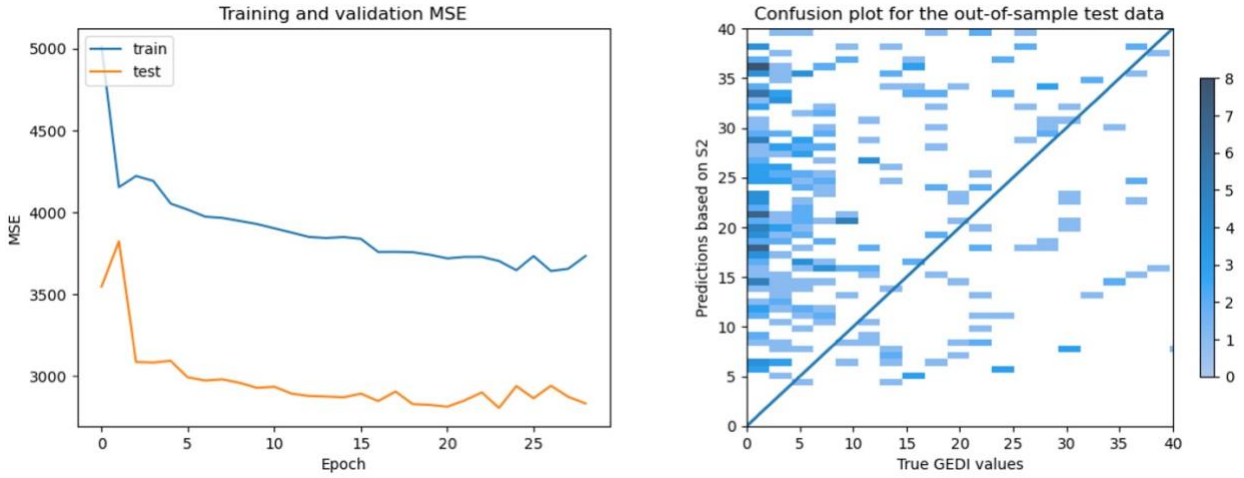

Figure 9: Learning curves and confusion plot for CNN (V1) with geographically separated training and test data

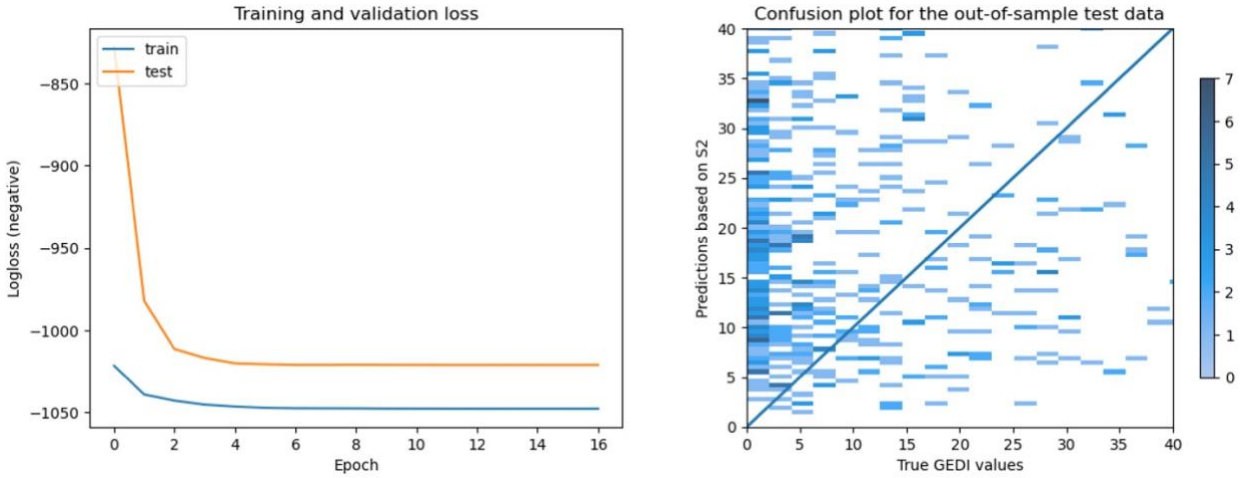

Figure 10: Learning curves and confusion plot for XGB (V1) with geographically separated training and test data

The results show that both models fare worse with the geographically separated data. As was to be expected, the CNN model catches up XGB in terms of MSE and the distribution on the confusion plot. This time the CNN has a lower MSE of 7071 vs 7799 for the XGB model. XGB's inability to extrapolate is something even the custom weights cannot solve.

4 Discussion and Conclusion

The results of this paper show that that optical images can be used to estimate biomass. Though maintaining some unevenly distributed errors, the models are able to predict biomass with some certainty. Nevertheless, all models (to varying degrees) struggle with overestimating areas with low biomass. This is a reflection of a very large bias in the training data. The data contains far more points with low GEDI estimations compared to points with large biomass estimations. Assigning a set of custom weights to the XGB loss function does address this issue for the XGB model and brings down the MSE from 3729 to 2675. This is a significant improvement over the baseline XGB model and the CNN model; however, there are model-related issues with XGB that make it less well-suited for this problem. The fact that the XGB model outperforms the CNN is likely due to the fact that the sample size (n=10,000)[19] is less than would be ideal for the CNN model. If the data was three orders of magnitude larger, the CNN would likely outperform XGB even with the weights applied.

Another limitation of the model is that it is limited by the information the optical (passive sensors) can capture. LIDAR is able to penetrate canopy-covers whereas passive sensing cannot. While the non-linearities of the CNN model do allow the model to pick up on relationships between the optical images and GEDI estimations that a human would likely not pick up on, there will always be some information lost. Particularly in the densest canopy covers, the limitations of the optical sensors will likely result in some error. Biomass estimation is useful for many applications; however, the temporal aspect of biomass accumulation is also important to track. Since the date ranges for the S2 images and the GEDI estimates line up the model will have some ability to make temporal predictions, though it is limited by the fact that the dates match up only within the range of two years. In order to improve the predictive capabilities of this model and reduce misclassifications a far larger data set would be a good starting point. Additionally, a more precise linking of dates would allow for temporal changes to be tracked better.

[19] Producing data orders of magnitude larger would likely require a different approach to the data gathering. At this time, the SentinelHub API is simply not fast enough to allow for the retrieval of millions of images in a reasonable timeframe.

Bibliography

Dubayah, R.O., J. Armston, S.P. Healey, Z. Yang, P.L. Patterson, S. Saarela, G. Stahl, L. Duncanson, and J.R. Kellner. "GEDI L4B Gridded Aboveground Biomass Density, Algorithm Theoretical Basis Document." ORNL Distributed Active Archive Center, 2022. https://doi.org/10.3334/ORNLDAAC/2017.

———. "GEDI L4B Gridded Aboveground Biomass Density, Version 2." ORNL Distributed Active Archive Center, 2022. https://doi.org/10.3334/ORNLDAAC/2017.

Eggleston, H. S., L. Buendia, K. Miwa, T. Ngara, and K. Tanabe. "2006 IPCC Guidelines for National Greenhouse Gas Inventories," July 1, 2006. https://www.osti.gov/etdeweb/biblio/20880391.

Fraser, Barbara. "Study Brings Satellite Data down to Earth for Biomass Tracking." CIFOR Forests News, October 4, 2022. https://forestsnews.cifor.org/79313/study-brings-satellite-data-down-to-earth-for-biomass-tracking-2?fnl=.

Gentine, Pierre. "EAEEE4000 Machine Learning for Environmental Engineering and Science Slides." Lecture Slides, Columbia University, Fall 2022.

IPCC. "Summary for Policymakers." In *Climate Change 2021: The Physical Science Basis, the Working Group I Contribution to the Sixth Assessment Report*, edited by Richard P Allan, Christophe Cassou, Deliang Chen, Annalisa Cherchi, L Connors, Francisco J Doblas-Reyes, Hervé Douville, et al. Cambridge University Press. In Press., 2021. https://www.ipcc.ch/report/sixth-assessment-report-working-group-i/.

Kumar, Lalit, and Onisimo Mutanga. "Remote Sensing of Above-Ground Biomass." *Remote Sensing* 9, no. 9 (September 2017): 935. https://doi.org/10.3390/rs9090935.

Land Product Validation Subgroup (Working Group on Calibration and Validation, Committee on Earth Observation Satellites). "Aboveground Woody Biomass Product Validation Good Practices Protocol," 2021. https://doi.org/10.5067/DOC/CEOSWGCV/LPV/AGB.001.

Lang, Nico, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. "A High-Resolution Canopy Height Model of the Earth." arXiv, April 13, 2022. https://doi.org/10.48550/arXiv.2204.08322.

"Sentinel Hub." Accessed December 24, 2022. https://www.sentinel-hub.com/.

The European Space Agency. "Sentinel-2 - Missions - Sentinel Online - Sentinel Online." Accessed December 24, 2022. https://sentinel.esa.int/web/sentinel/missions/sentinel-2.

Sinergise. "Sentinel-2 L2A." Sentinel Hub by Sinergise. Accessed December 24, 2022. https://docs.sentinel-hub.com/api/latest/data/sentinel-2-l2a/.