Supplementary Material

Straightforward Model-Based Approach Using Only Field Data and Open-Source Maps to Improve Carbon Stock Estimates for REDD+ Projects

Leo Eiti Haneda[a*], Danilo Roberti Alves de Almeida[b], Renan Akio Kamimura[b], Silvio H. M. Gomes[b], Arthur K. Sanchez[b,c], Bruna Pereira de Azevedo[b], Cléber Rodrigo de Souza[b,d] , Rafaela Pereira Naves[b], Denis Valle[a]

[a] School of Forest, Fisheries and Geomatics Sciences, University of Florida, Gainesville, FL, US

[b] BRCarbon, Piracicaba, SP, BR

[c] Department of Forest Sciences, "Luiz de Queiroz" College of Agriculture, University of São Paulo, Piracicaba, SP, BR

[d] Forest Sciences Department, Federal University of Lavras, MG, BR

*Corresponding author: Tel: +1 (352) 745-9727; Email: hanedaleo@ufl.edu

Models used in the study:

Linear Regression Model (Lm) – We used the *lm()* function from the basic stats package implemented in R to fit simple linear regressions models [1].

Random Forest (RF) – The RF models were fitted using the function *randomForest()* from the *randomForest* R package, using all parameters set to the default settings [2]. Random Forest is a powerful machine learning algorithm belonging to the ensemble learning family, and it is based on decision trees. Each tree is trained using bootstrapped samples and randomly selected predictor variables and the predictions of these multiple decisions trees are combined to produce a more accurate final prediction[3].

Spatial Nearest Neighbor Gaussian Processes (spNNGP) – We used the *spNNGP* R package [4]. We fit this model using the *spNNGP()* function, applying the "latent" method with the default settings. This package implements spatial regression models for Gaussian and non-Gaussian univariate outcomes and the software provides an efficient algorithm to fit spatially correlated data.

Generalized Additive Model (GAM) – To fit the GAM models, we used the *gam()* functions from the *mgcv* R package [5], applying the "REML" method. GAM is a flexible class of generalized linear model (GLM) that involves a sum of smooth functions of covariates and allows for non-linear relationships [6]. The advantage of this model is the ability to account for spatial correlation in the data by adding 2D splines of the spatial coordinates.

Large-scale Maps used in the study:

Name	Publisher	Spatial Resolution	Data Inputs	Models/Algorithm	References / Additional Information
1. CCI 2020 Global Aboveground Biomass Map	ESA	100 m	Field plots, GEDI, Sentinel-1 SAR C-band, and ALOS-2 PALSAR-2 L-band	BIOMASAR algorithm (inverted Water-Cloud-like model)	[7,8]
2. Amazon Aboveground Biomass Density Map	Ometto et al. [9]	250 m	Field plots, ALS, ALOS-2 PALSAR-2, MODIS vegetation indeces, SRTM, and TRMM data	Random Forest regression algorithm	[9]
3. Global Canopy Height Map	Meta and WRI	1 m	Field plots, ALS, Maxar Vivid2 mosaic imagery (WorldView-2, WorldView-3, Quickbird II), and GEDI	DiNOv2 (Self-Supervised Learning model)	[10,11]
4. Global Canopy Height Map	Lang et al. [12]	10 m	Field plots, GEDI, and Sentinel-2	CNN architecture	[12]

Table 1. Open-source large-scale maps evaluated in the study. Note: CCI = Climate Change Initiative, ESA = European Space Agency, GEDI = Global Ecosystem Dynamics Investigation, SAR = Synthetic Aperture Radar, ALOS = Advanced Land Observing Satellite, PALSAR = Phased Array L-band Synthetic Aperture Radar, ALS = Airborne Laser Scanning, MODIS = Moderate-Resolution Imaging Spectroradiometer, SRTM = Shuttle Radar Topography Mission, TRMM = Tropical Rainfall Measuring Mission, WRI = World Resources Institute, CNN = Convolutional Neural Network.

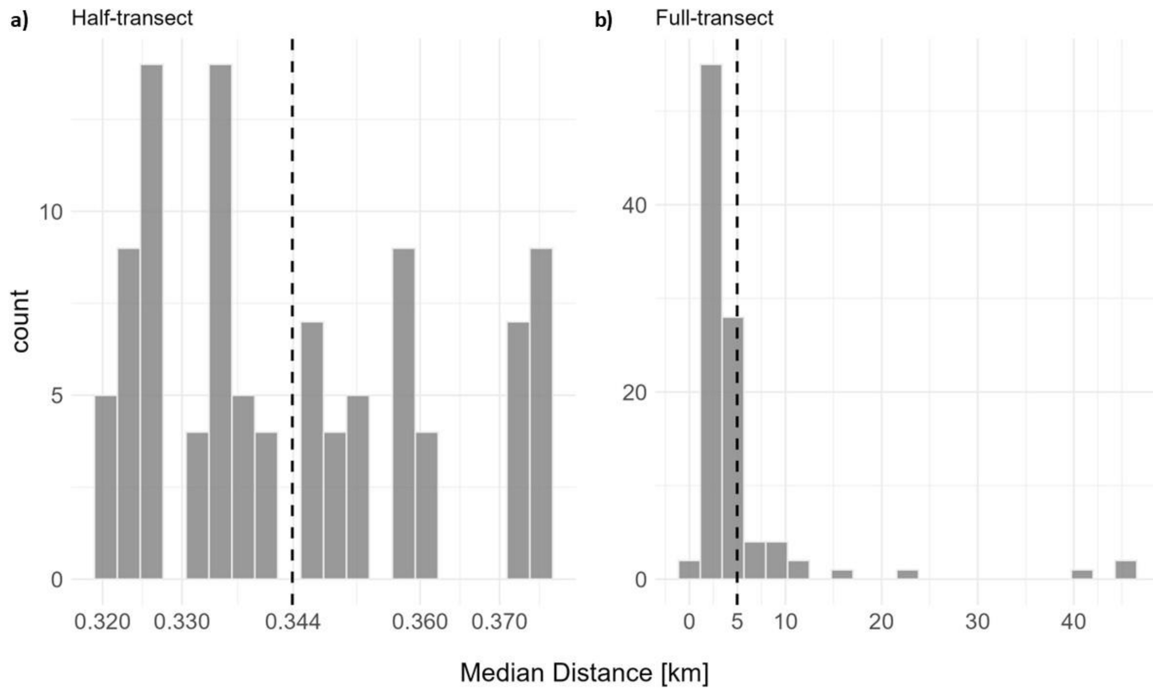

Figure S1. Histogram of median distances between training and testing subsets in each iteration for the half-transect (**a**) and full-transect (**b**) validation schemes. The dashed line indicates the mean value in each histogram.

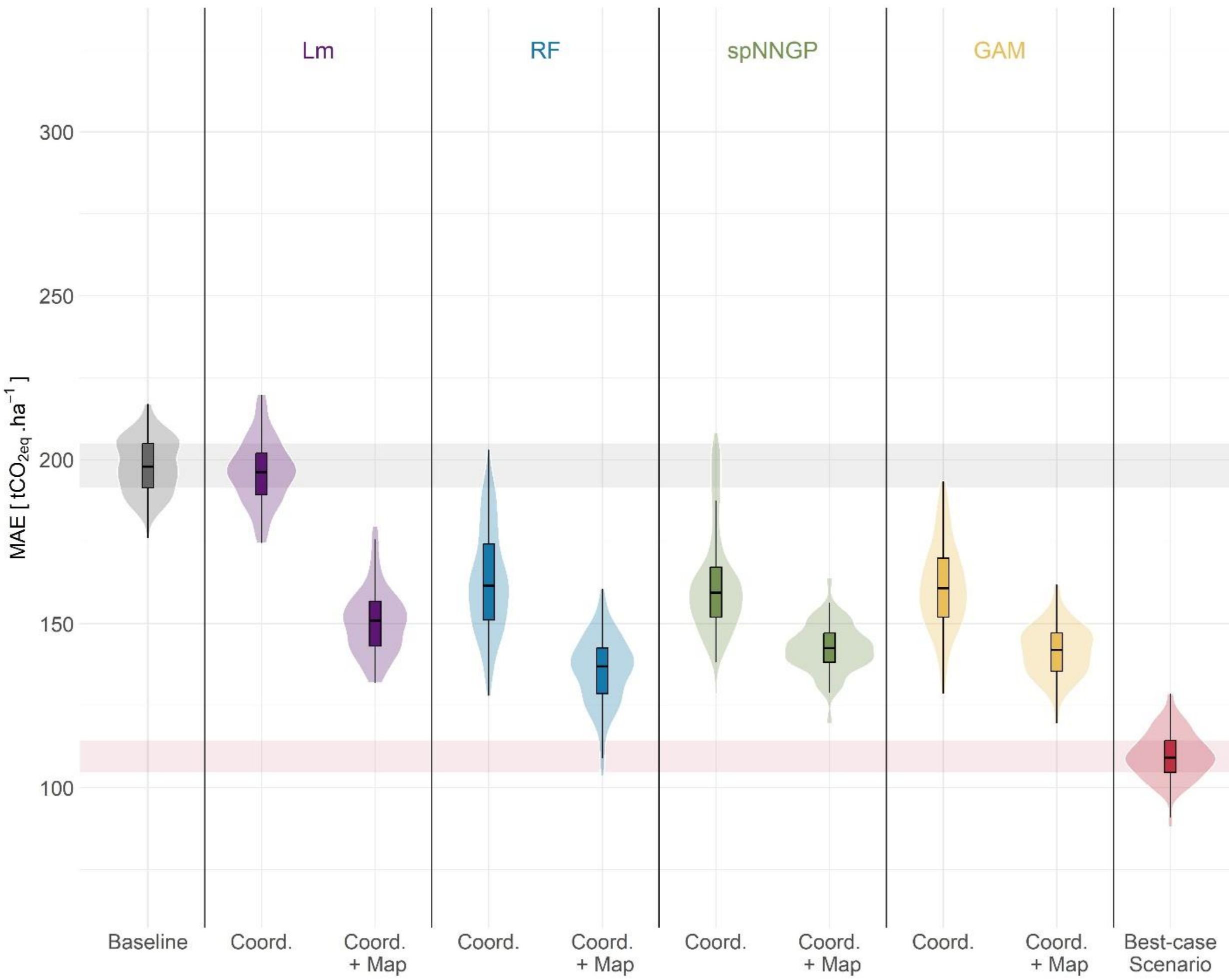

Figure S2. Half-Transect validation results for the different models and predictor variables combinations based on mean absolute error (MAE). The grey and red zones indicate the range of values for the baseline and best-case scenarios, respectively.

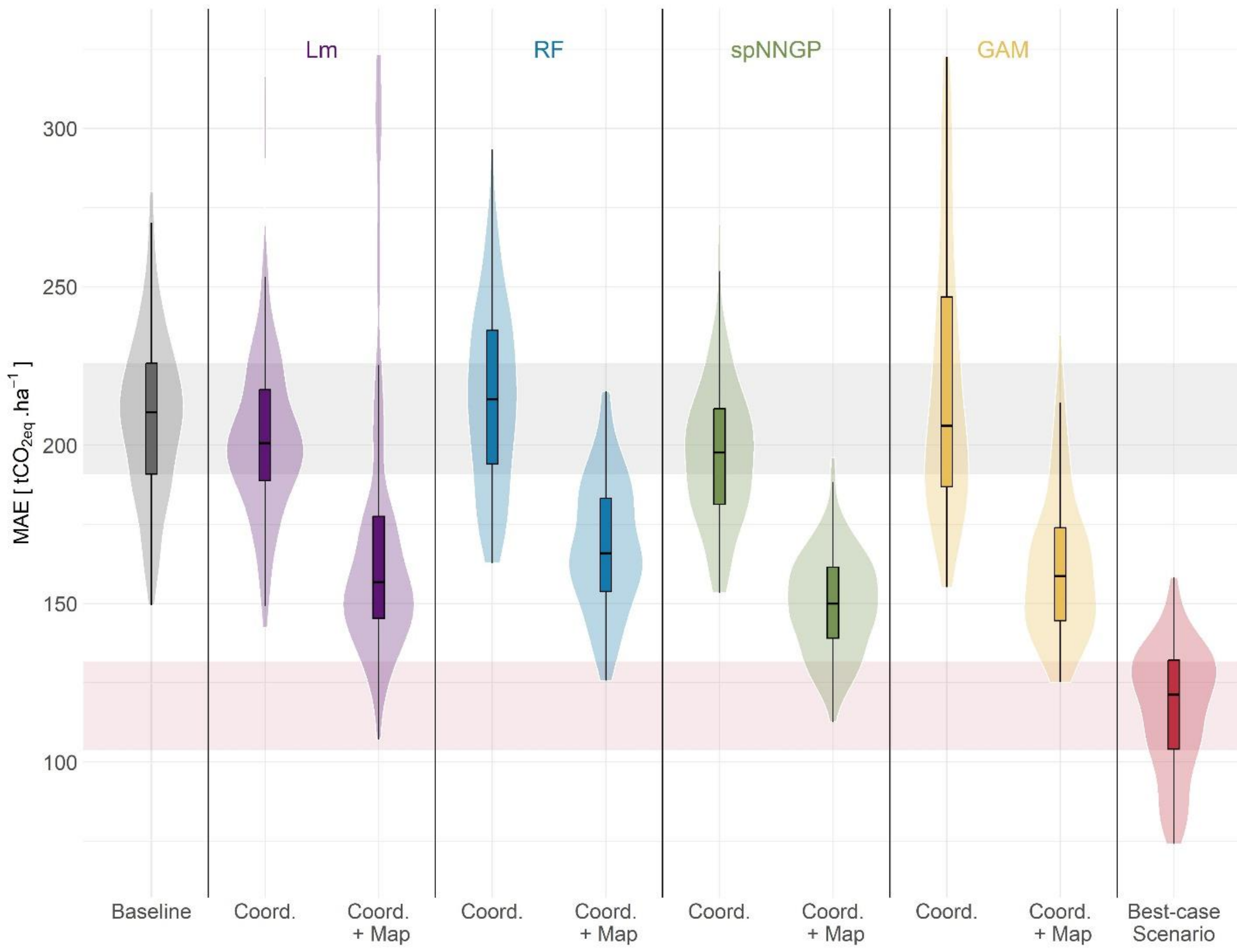

Figure S3. Full-Transect validation results for the different models and predictor variables combinations based on mean absolute error (MAE). The grey and red zones indicate the range of values for the baseline and best-case scenarios, respectively.

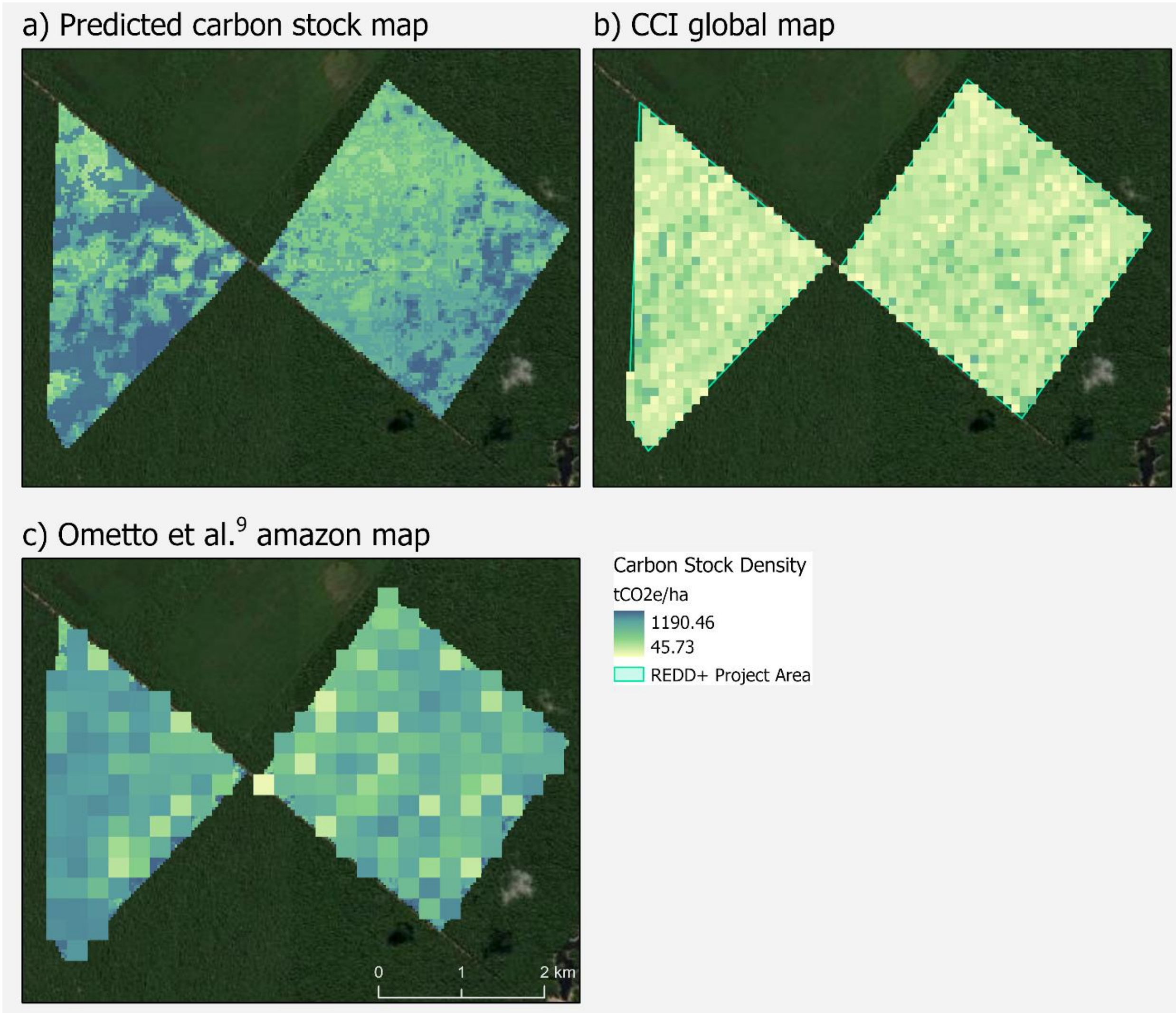

Figure S4. Comparison between the predicted carbon stock map derived using our methodology and open-source biomass maps as references. We converted AGB to Carbon Stock Density (CSD) in **b** and **c** in accordance with IPCC guidelines.

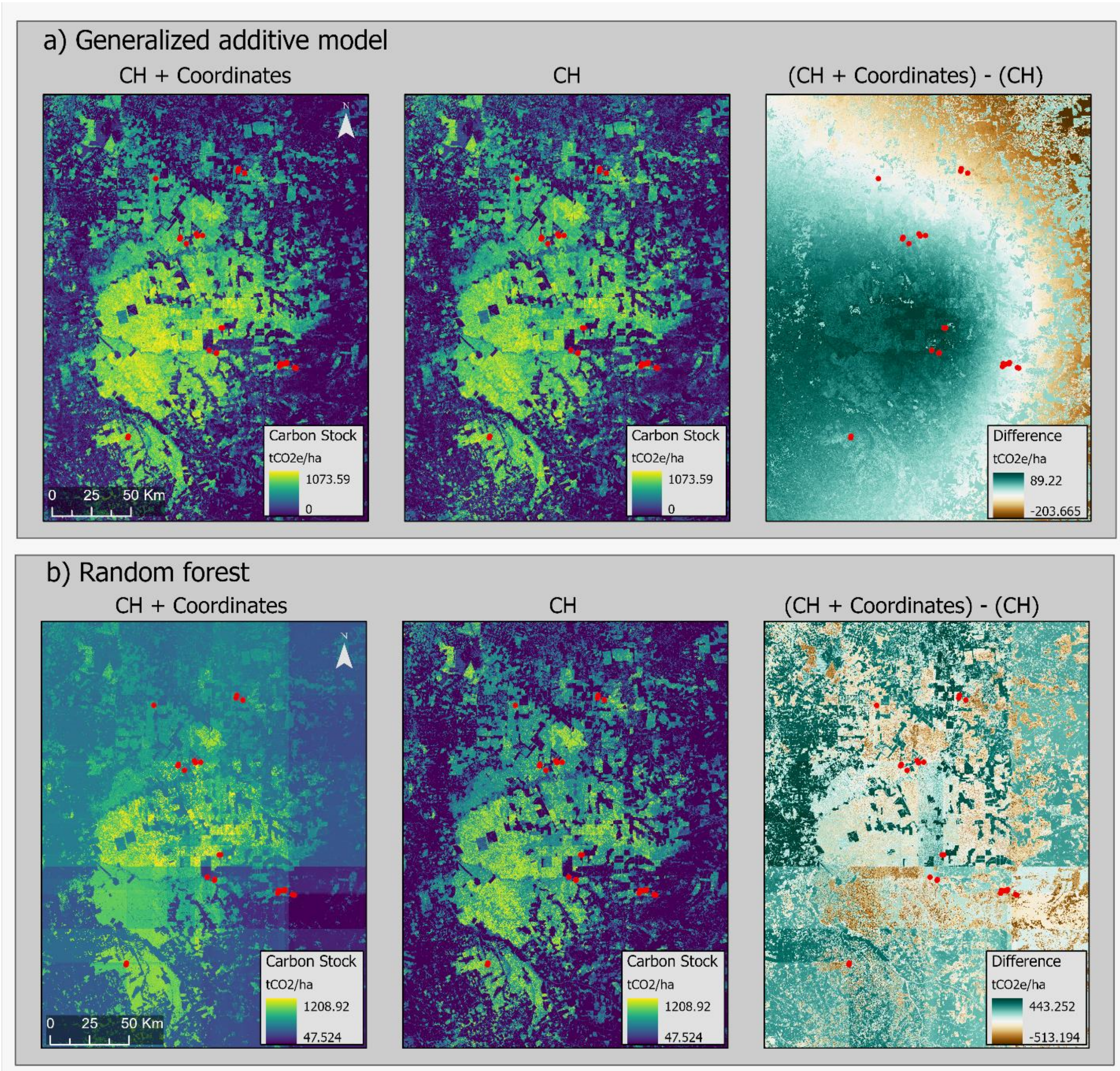

Figure S5. Effect of accounting for spatial correlation on regional carbon stock maps. Left panels show model predictions using spatial coordinates and canopy height as covariates. Middle panels show model predictions using only canopy height as covariate. Finally, right panels show difference between first and second map (left minus middle panels). Top panels show results based on Generalized Additive Model while bottom panels show results based on Random Forest.

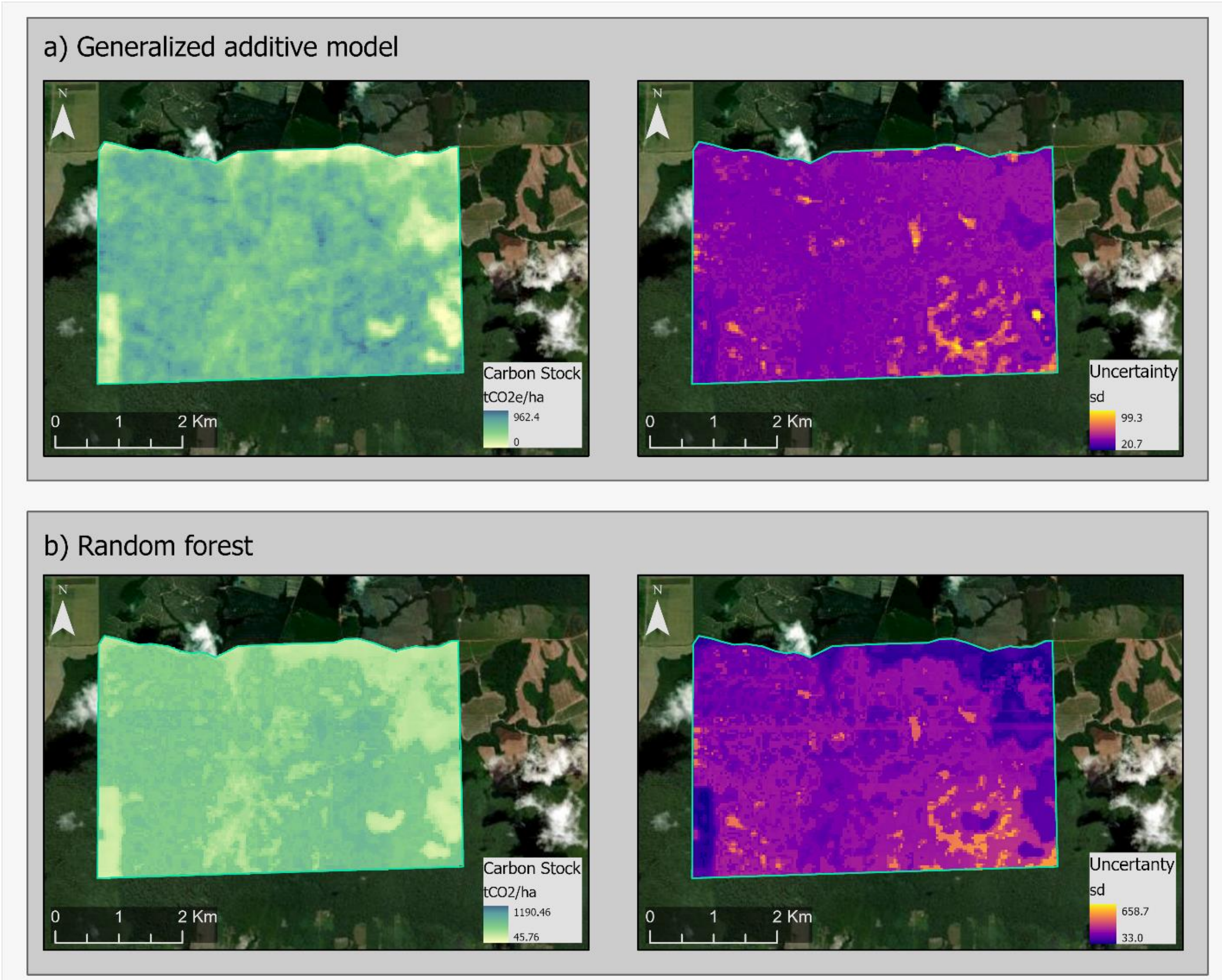

Figure S6. Examples of uncertainty map for a REDD+ project area using GAM (**a**) and Random Forest (**b**). Left: Carbon stock map using spatial coordinates and canopy height as covariates. Right: Uncertainty maps of model predictions. Sd stands for standard deviation.

References

1. R Core Team. R: A Language and Environment for Statistical Computing. Preprint at https://www.R-project.org/ (2023).
2. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* Vol. 2/3, (2002).
3. Breiman, L., Cutler, A., Liaw, A. & Wiener, M. randomForest: Breiman and Cutlers Random Forests for Classification and Regression. *CRAN: Contributed Packages* Preprint at https://doi.org/10.32614/CRAN.package.randomForest (2002).
4. Finley, A. O., Datta, A. & Banerjee, S. spNNGP R Package for Nearest Neighbor Gaussian Process Models. *J Stat Softw* 103, (2022).
5. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Stat Soc Series B Stat Methodol* 73, 3–36 (2011).
6. Wood, S. N. *Generalized Additive Models: An Introduction with R*. (Chapman and Hall/CRC, 2017). doi:10.1201/9781315370279.
7. Santoro, M. & Cartus, O. ESA Biomass Climate Change Initiative (Biomass_cci): Global datasets of forest above-ground biomass for the years 2010, 2017, 2018, 2019 and 2020, v4. *NERC EDS Centre for Environmental Data Analysis* https://doi.org/10.5285/af60720c1e404a9e9d2c145d2b2ead4e (2023).
8. Santoro, M. *et al.* The BIOMASAR algorithm: An approach for retrieval of forest growing stock volume using stacks of multi-temporal SAR data. *Proceedings of ESA Living Planet Symposium* (2010).
9. Ometto, J. P. *et al.* A biomass map of the Brazilian Amazon from multisource remote sensing. *Sci Data* 10, 668 (2023).
10. Oquab, M. *et al.* DINOv2: Learning Robust Visual Features without Supervision. *ArXiv* https://doi.org/10.48550/arxiv.2304.07193 (2023) doi:10.48550/arxiv.2304.07193.
11. Tolan, J. *et al.* Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sens Environ* 300, 113888 (2024).
12. Lang, N., Jetz, W., Schindler, K. & Wegner, J. D. A high-resolution canopy height model of the Earth. *Nat Ecol Evol* 7, 1778–1789 (2023).