ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

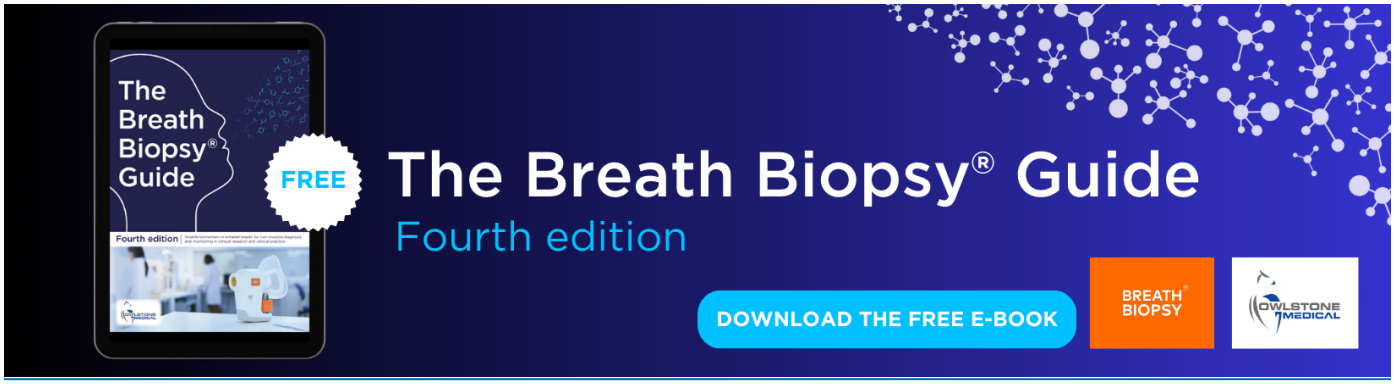
Using GEDI as training data for an ongoing mapping of landscape-scale dynamics of the plant area index

To cite this article: Alice Ziegler *et al* 2023 *Environ. Res. Lett.* **18** 075003

View the article online for updates and enhancements.

You may also like

- Inferring alpha, beta, and gamma plant diversity across biomes with GEDI spaceborne lidar
C R Hakkenberg, J W Atkins, J F Brodie et al.
- The use of GEDI canopy structure for explaining variation in tree species richness in natural forests
Suzanne M Marselis, Petr Keil, Jonathan M Chase et al.
- GEDI launches a new era of biomass inference from space
Ralph Dubayah, John Armston, Sean P Healey et al.



This content was downloaded from IP address 137.248.1.31 on 18/01/2024 at 09:13

IOP Publishing Environ. Res. Lett. 18 (2023) 075003 https://doi.org/10.1088/1748-9326/acde8f

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
17 March 2021

REVISED
12 June 2023

ACCEPTED FOR PUBLICATION
15 June 2023

PUBLISHED
27 June 2023

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

LETTER

Using GEDI as training data for an ongoing mapping of landscape-scale dynamics of the plant area index

Alice Ziegler[1,*], Johannes Heisig[2], Marvin Ludwig[3], Chris Reudenbach[4], Hanna Meyer[3] and Thomas Nauss[1]

[1] Environmental Informatics, Faculty of Geography, University of Marburg, Deutschhausstraße 12, 35032 Marburg, Germany
[2] Institute for Geoinformatics, University of Muenster, Heisenbergstraße 2, 48149 Münster, Germany
[3] Institute Landscape Ecology, University of Muenster, Heisenbergstraße 2, 48149 Münster, Germany
[4] Environmental Modeling, Faculty of Geography, University of Marburg, Deutschhausstraße 10, 35032 Marburg, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: alice.ziegler@geo.uni-marburg.de

Keywords: random forest, LAI, leaf area index, spatial modeling, maschine learning, environmental monitoring, phenology

Abstract

Leaf or plant area index (LAI, PAI) information is frequently used to describe vegetation structure in environmental science. While field measurements are time-consuming and do not scale to landscapes, model-based air- or space-borne remote-sensing methods have been used for many years for area-wide monitoring. As of 2019, NASA's Global Ecosystem Dynamics Investigation (GEDI) mission delivers a point-based LAI product with 25 m footprints and periodical repetition. This opens up new possibilities in integrating GEDI as frequently generated training samples with high resolution (spectral) sensors. However, the foreseeable duration of the system installed on the ISS is limited. In this study we want to test the potential of GEDI for regional comprehensive LAI estimations throughout the year with a focus on its usability beyond the lifespan of the GEDI mission. We study the landscape of Hesse, Germany, with its pronounced seasonal changes. Assuming a relationship between GEDI's PAI and Sentinel-1 and -2 data, we used a Random Forest approach together with spatial variable selection to make predictions for new Sentinel scenes. The model was trained with two years of GEDI PAI data and validated against a third year to provide a robust and temporally independent model validation. This ensures the applicability of the validation for years outside the training period, reaching a total RMSE of 1.12. Predictions for the test year showed the expected seasonal and spatial patterns indicated by RMSE values ranging between 0.75 and 1.44, depending on the land cover class. The overall prediction performance shows good agreement with the test data set of the independent year which supports our assumption that the usage of GEDI's PAI beyond the mission lifespan is feasible for regional studies.

1. Introduction

Leaf area index (LAI) describes the structure of vegetation as the ratio of leaf area to ground area. LAI is a highly relevant variable for interactions between vegetation and atmosphere and is therefore one of the proclaimed Essential Climate Variables (GCOS 2021). The plant area index (PAI) is defined as half of the total plant area per unit ground surface. Compared to the LAI, not only leaves, but all above-ground plant components such as branches and trunks are considered. However, the PAI is closely related to the LAI (Myneni *et al* 2001, Weiss *et al* 2007, Feret *et al* 2008, Tang *et al* 2012). On a global scale, carbon flux and evapotranspiration are massively influenced by LAI. At the regional scale, in addition to its use in models for carbon assessment e.g. (Tharammal *et al* 2019), the importance of LAI is, for example, evident in run-off models as it affects not only evapotranspiration rates but also immediate water retention (Tesemma *et al* 2015, Seo and Kim 2021, Huang *et al* 2022). Since LAI varies greatly due to seasonality, the use of a temporally as well as spatially (Huang *et al* 2022) detailed data set could help improve dynamic modeling of such variables.

Various field methods and techniques exist for the direct (destructive) measurement and the indirect estimation or inverse modeling of the LAI (Zheng

© 2023 The Author(s). Published by IOP Publishing Ltd

and Moskal 2009, Fang *et al* 2019). However, direct or field-based methods in general are time-consuming and therefore do not scale over time and space. In contrast, air-borne light detection and ranging (LiDAR) missions provide high-resolution wall-to-wall data on vegetation structure at landscape level and can be used for spatial mapping of LAI (Yan *et al* 2019, Wang and Fang 2020). Nonetheless, they have limitations with respect to temporal repetition, which usually does not allow for spatial time series that are dense enough to derive phenological information. Global satellite-based products, e.g. from MODIS (Myneni *et al* 2001, Qiao *et al* 2019) or Sentinel-3 (Fuster *et al* 2020) fill this gap by providing regular repetitions, but the resolutions of typically 300 m to 1 km are too coarse for small-scale differentiated studies. To bridge this gap between current global space-borne products and air-borne or field-based missions, the utilization of higher resolution radar and optical sensor satellites, e.g. from Sentinel-1 or -2 data, is promising (Frampton *et al* 2013, Baghdadi *et al* 2016, Pasqualotto *et al* 2019, Wang *et al* 2019, Kganyago *et al* 2020, Luo *et al* 2020, Padalia *et al* 2020). Yet, this requires extensive ground truth data and sophisticated modeling strategies, to link optical and radar data to the response variable—the LAI. This leads back to the point that field observations are not sufficiently comprehensive in the spatial and temporal domain and hence do not provide a sufficient baseline for model training. Bringing LiDAR into space with NASA's new Global Ecosystem Dynamics Investigation (GEDI) mission in December 2018 was a big step towards almost global, space-borne, and direct observation of vegetation structure which may provide training and testing samples with a much higher temporal repetition rate. Since January 2019 data sets are available and studies confirm the high potential for ecosystem monitoring (Boucher *et al* 2020, Healey *et al* 2020, Marselis *et al* 2020, Di Tommaso *et al* 2021, Kacic *et al* 2021, Potapov *et al* 2021, Rishmawi *et al* 2021, 2022, Wang *et al* 2022, Xi *et al* 2022).

A Level 2B standard product of GEDI is the PAI that is closely related to the LAI (see section 2.2.1). Version 2 of the PAI product provides information for footprints with 25 m in diameter with a spacing of 60 m along track and 600 m across track (Dubayah *et al* 2020). To derive wall-to-wall products from GEDI's PAI, ESA's Sentinel-1 and Sentinel-2 systems are promising candidates, as vegetation structure interferes with radar and optical wavelengths, and since both sensors come with a high spatial and temporal resolution. The radar observations of Sentinel-1 capture vertical vegetation heterogeneity similar to LiDAR observations (Bae *et al* 2019) and the multi-spectral scanner observations of Sentinel-2 are indicators for plant physiology, vegetation type and biomass. Previous studies have shown a high potential to estimate LAI from a combination of different spectral bands using non-linear models (Verrelst *et al* 2015, Baghdadi *et al* 2016, Korhonen *et al* 2017, Wang *et al* 2019, Jiang *et al* 2020, Luo *et al* 2020).

The aim of this study is to use the PAI as estimated by GEDI to produce wall-to-wall maps by an integration of Sentinel optical and radar data. The integration of GEDI's PAI and Sentinel-1/-2 data was already proven to be feasible (Di Tommaso *et al* 2021, Kacic *et al* 2021, Rishmawi *et al* 2021, 2022). However, previous studies that aimed at wall-to-wall mapping of GEDI-derived variables used temporally aggregation, i.e. long-term means of such variables (Healey *et al* 2020, Potapov *et al* 2020, Chen *et al* 2021, Dorado-Roda *et al* 2021, Khati *et al* 2021, Verhelst *et al* 2021, Francini *et al* 2022, Shendryk 2022). This is certainly suitable for a detection of large-scale spatial patterns, however, does not support seasonality and is hence not suitable for studies that require the consideration of phenology. The potential of learning the seasonal dynamics by using the different overpasses of the GEDI with the corresponding optical and radar data from Sentinel is, to our best knowledge, not analyzed yet. In this study we test the potential of matching GEDI point data to the temporally closest Sentinel scene in a machine learning approach, with the aim to derive wall-to-wall predictions of the PAI with a temporal resolution that is in accordance with the availability of the frequent Sentinel scenes. Motivation for this study comes from the limited duration of the GEDI mission onboard the ISS. Therefore it is of particular importance to apply and test the trained model beyond its training period. Hence, in this study, we use two of the three years of GEDI data to train a model and we validate the performance with the remaining year. The German state Hesse was selected as a study area because of its heterogeneous landscape of forests, pastures and cultivated land. Since we expect land cover to cause differences in model performance, results are interpreted by taking the different types into account.

2. Methods

2.1. Study area

The state of Hesse, Germany (about 21 000 km^2) was used as the study area (see figure 1). The area with low mountain ranges and a temperate climate with pronounced seasonal changes is composed of 26% non-irrigated arable land, 25% broad-leaved forest, 20% pastures, 13% coniferous forest and 3% mixed forest according to the Coordination of information on the environment (CORINE) land cover inventory 2018 (European Union, Copernicus Land Monitoring System, 2018). The remaining area is mostly covered by urban areas and some smaller patches of other land cover classes that will not be considered in this study.

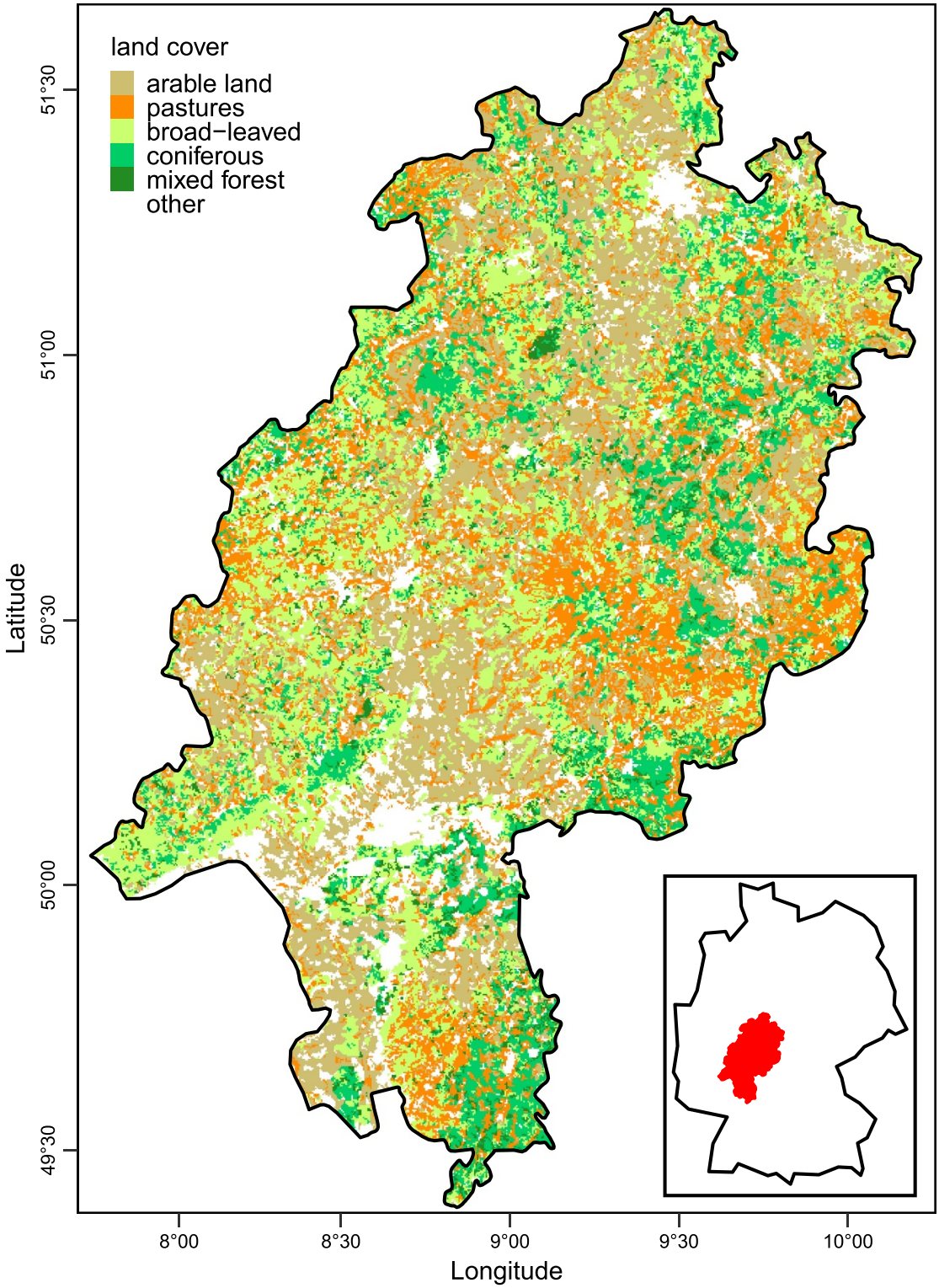


Figure 1. CORINE land cover classes for the study area Hesse. Small sketch in the bottom-right corner shows the location of Hesse within Germany.

2.2. Data sources

Data pre-processing was mainly executed in Google Earth Engine (GEE) to handle the large data volume (Gorelick *et al* 2017). All steps of model training and evaluation were performed in *R* (R Core Team 2022). For the availability of scripts and data sets see the data availability statement (Ziegler and Ludwig 2023).

2.2.1. PAI GEDI data

The GEDI mission was operational from March 2019 to March 2023 and is tentatively scheduled to provide

additional data beginning in the fall of 2024. All orbits that intersected the study area within the study period were identified using the rGEDI package in *R* (Silva *et al* 2021). The level 2B version 2 PAI product was obtained within GEE and contains footprints with a diameter of 25 m for eight parallel tracks spaced 600 m across track and 60 m along track (Dubayah *et al* 2020). For our study area it offers about one daily overpass but no complete spatial coverage or overlap (Dubayah *et al* 2020). Of the available 1441 GEDI overflights from April 2019 until December 2021, 1347 potential overflights remained after excluding footprints with a low quality flag and GEDI's sensitivity value below 0.9 in the Level 2B PAI product (Tang *et al* 2019).

2.2.2. Sentinel-1 data

To provide spatially continuous predictors for PAI, the ground-range-detected high-resolution product of Sentinel-1's C-Band radar, retrieved directly within GEE, was used for the study. Both polarization modes (VV: vertically transmitted and received, VH: vertically transmitted and horizontally received) with a resolution of 10 m and their difference and ratio (VV-VH, VV/VH) were used as potential predictor variables since these metrics showed to be sensitive to seasonal changes in forested areas (Frison *et al* 2018). For the selection process of appropriate scenes and the matching with GEDI data, see section 2.3.1.

2.2.3. Sentinel-2 data

Sentinel-2 Level 2A multi-spectral data, retrieved directly within GEE, provided the second source of potential predictor variables for the PAI. All available spectral bands except band 10, which lies within the atmospheric absorption bands of water vapor and carbon dioxide, were selected as potential predictors in their original resolution. The normalized difference vegetation index (NDVI), enhanced vegetation index (EVI) and the inverted red-edge chlorophyll index, that copes well with the problem of oversaturation in LAI products, were computed (Frampton *et al* 2013). For the selection process of appropriate scenes and the matching with GEDI data, see section 2.3.1.

2.3. Processing Methods

2.3.1. Temporal and spatial matching of GEDI data with Sentinel-1/-2

Temporally matching Sentinel data were queried for each of the 1441 available GEDI overpasses. We allowed a temporal difference of +/− 3 days for Sentinel-1 and +/− 5 days for Sentinel-2 and a maximum cloud cover of 50% per scene. If more than one scene was available within this window, we chose the temporally closest cloud-free pixel to the GEDI acquisition date. All pixels selected according to this procedure were combined into a mosaic and used in the following for the corresponding GEDI recording. If no adequate Sentinel data were available within this time window, the GEDI overpass was not considered for further analysis. For each of the 25 m footprints, all intersecting pixels were extracted from the mosaic and used as predictor variables for the model.

2.3.2. Data quality and sampling strategy

The availability of sufficiently large data sets allowed generous filtering to use only high quality data points, especially because the GEDI location error adds additional uncertainty. Therefore, GEDI footprints were excluded if located within a 1000 m buffer around pixels marked as clouds or cloud shadows or flagged as defective by Sentinel's scene classification band. If footprints were flagged in the GEDI quality band or covered mixed land cover classes according to the CORINE inventory they were excluded from the study. Additionally, only the five main land cover classes were taken into consideration (see section 2.1). Outliers were rigorously eliminated from Sentinel and PAI data by excluding the upper and lower 0.1% of the data points. This temporal and spatial matching as well as sub-sampling resulted in 7132 148 valid observations from 1327 GEDI overpasses between April 2019 and December 2021. Due to the massive amount of data points we randomly sampled 150 000 of those high quality points for model training (127 449 points) and testing (22 551 points, see figure 2). The number of samples for the training data set across land cover classes and month is visible in figure 2 and roughly follows the distribution corresponding to the proportions of land cover classes (see section 2.1) and prevailing weather conditions of the seasons.

2.3.3. Model training

In sum 16 predictors formed the set of training data: two from Sentinel-1 (see section 2.2.2); nine Sentinel-2 channels and five derived indices. A random forest machine learning approach was used to model PAI with the *R* packages caret (Kuhn 2008), ranger (Wright and Ziegler 2017) and CAST (Meyer 2020). During model tuning a spatial cross-validation (CV) was applied to evaluate the potential of models to predict GEDI PAI for new spatial areas. Therefore, we divided the data into 20 folds by keeping footprints from one orbit placed in the same fold, to ensure spatial and temporal independence between folds. For training, we only took data from 2019 and 2020 and used the 2021 data to assess the ability of the trained model to make predictions beyond the training phase. During model tuning, a spatial forward feature selection as explained in Meyer *et al* (2018) and Meyer *et al* (2019a) was applied: from all 18 potential predictor variables, only those that led to the lowest spatial CV error root-mean-square error (RMSE) were selected for the model training. The models were trained with 50 trees and the hyperparameter mtry was tuned with three different numbers of variables included in the respective training iteration step (2,

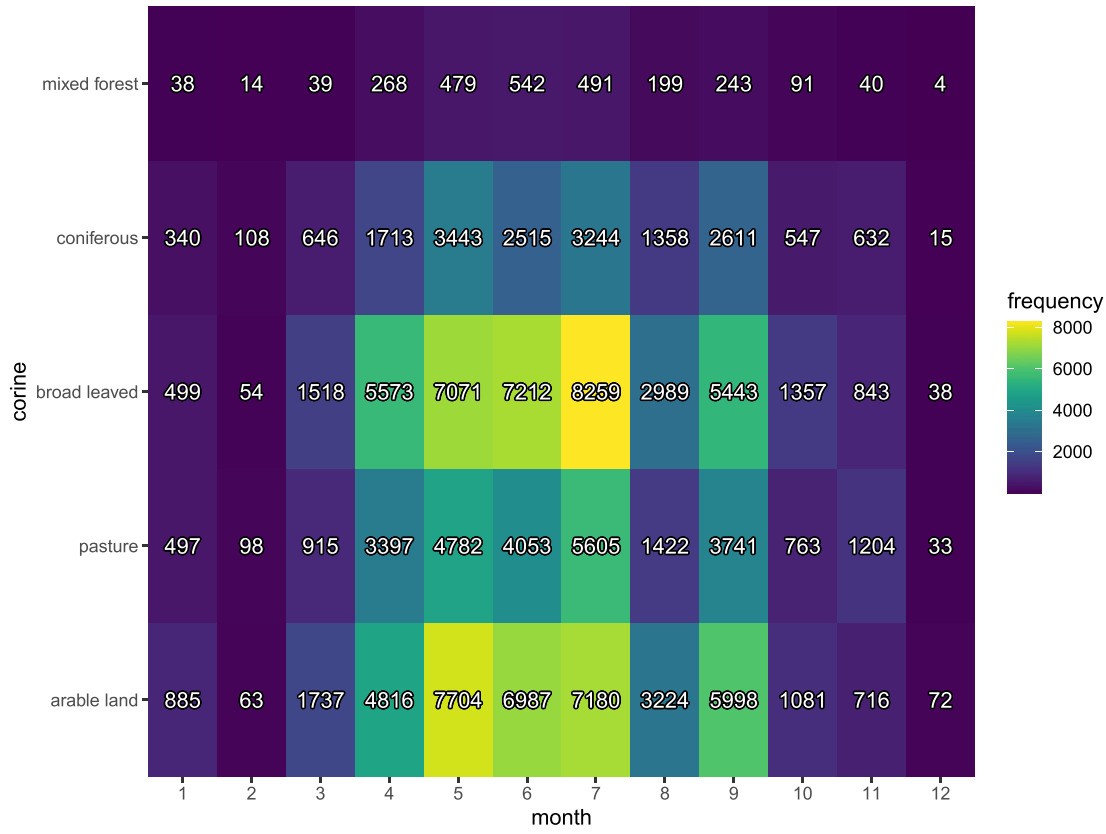


Figure 2. Number of available PAI observations in training data by month and land cover class.

6, 11). After tuning, model performance was quantified with the remaining 2021 data by computing R^2 and RMSE. Since the land use classes differ strongly in their variance, the RMSE was additionally normalized with the standard deviation (RMSE/sd) to allow a direct comparison of the performances between the different classes. To visually interpret the spatial patterns of the predictions, we applied the model to monthly composites of the predictors for the year 2021.

3. Results

3.1. Comparison of the temporal phenology dynamics of GEDI PAI and Sentinel-2 NDVI

To assess the plausibility of the GEDI PAI dynamics compared to the well-established NDVI, we first assessed the time series of both data sets. The temporal development of PAI from GEDI and NDVI from Sentinel-2 show clear patterns of vegetation phenology. While the NDVI reflects the phenology of vegetation by spectral properties responding to green biomass, the PAI reflects phenology by a change in the structure. The temporal dynamics in both, PAI and NDVI, reflect general differences especially during the summer months between arable land and pastures (lower values) and different forest types (higher values), but feature a large within-class variability (figure 3). The forest bud burst and leaf growth is well represented in the NDVI between February and May and also clearly visible in the PAI. The three forest types show different seasonal patterns in the NDVI, which also corresponds to the patterns in PAI. For pastures, the similarity between NDVI and PAI dynamics is less distinctly compared to the forest classes, as expected. However, the same tendencies can be observed considering an increasing variability during the summer months. The most obvious difference in PAI compared to what is reflected by the NDVI is observed for arable land. Here the NDVI shows clear seasonal patterns, which are not reflected by the median monthly PAI.

3.2. Assessment of the trained model

The spatial forward feature selection revealed that eleven out of 16 predictors were useful for spatial predictions of the PAI (figure 4). The best seven predictors include two visible bands (bands 3 and 4), two Sentinel-1 bands (VV and VH), two near-infrared bands from the red edge spectrum (bands 5 and 7) and one vegetation index (EVI). Both Sentinel-1 indices, and the two vegetation indices from Sentinel-2 data as well as the short wave infrared band (band 9) could not improve the model further.

The cross validation error for the 2019 and 2020 data set reached a R^2 value of 0.45 (RMSE: 1.20, RMSE/sd: 0.83). Based on the independent validation data set from 2021, the model reached an overall R^2 performance of 0.40 (RMSE: 1.12, RMSE/sd:

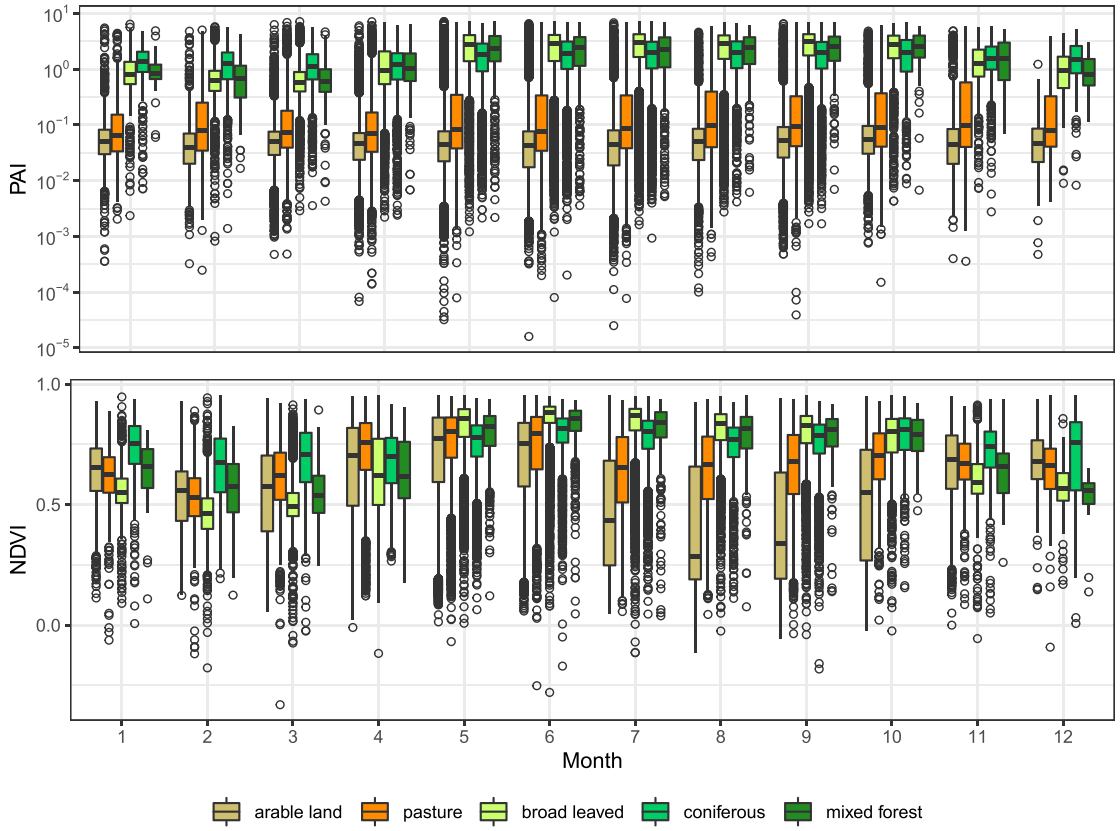


Figure 3. Temporal variability of the GEDI's PAI (top) and Sentinel's NDVI (bottom) per land cover class. The NDVI values are extracted from the same locations as the PAI and include a total of 150 000 points across the years 2019–2021.

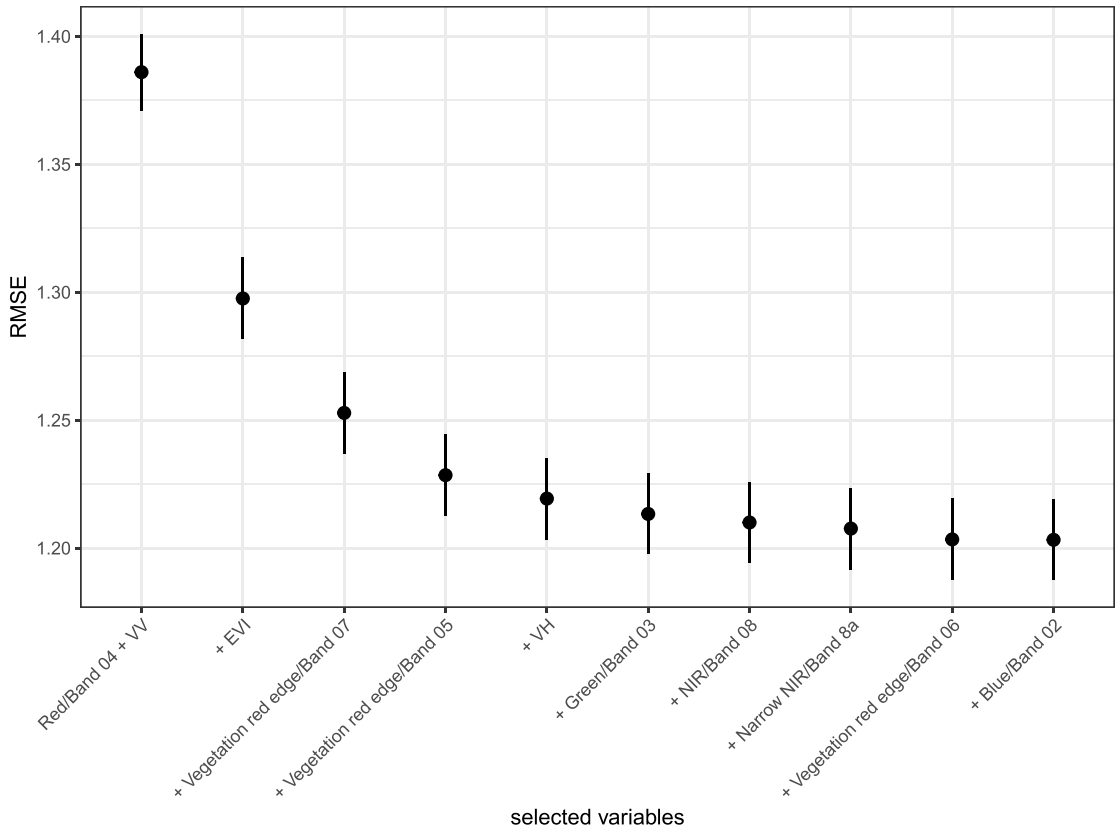


Figure 4. Ranking by RMSE of the predictor variables that were chosen in the feature selection to be included in the model.

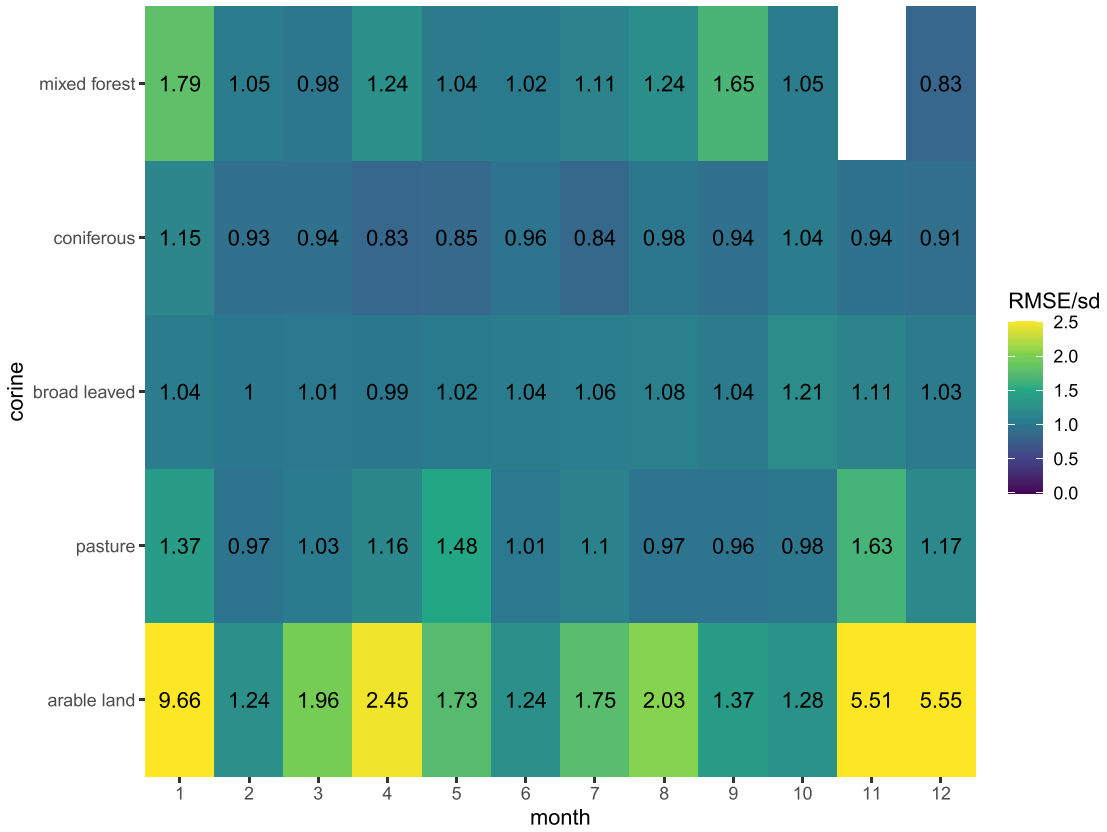


Figure 5. Assessment of prediction errors (RMSE/sd) for the 2021 validation data by month and land cover class.

0.78). The difference between the CV performance and the external validation statistics can be regarded as an indicator of the models potential to be applied beyond the training period. A large difference would indicate that the model performs significantly better within the training period. In our case, the CV error (gray) agrees well with the external test data set when validated across the land cover classes.

Figure 5 shows a heat map of the RMSE/sd for the independent validation by land cover class and month. Arable land performed worse than forest classes, overall and for each individual month except for September. In the winter months they fall far behind other classes with RMSE/sd above 5 (see figure 5). Forested classes follow a seasonal pattern with better performance in the summer months. The ranking of the performances of the forested classes alternates (see figure 5). In most months, coniferous forest has the best performance, despite an overall worse performance than mixed or broad-leaved forest (see table 1). The monthly breakdown of the external validation (see figure 6) demonstrates that all RMSE/sd values are below 1 and reveals a seasonal pattern with a better performance during the summer months. Regarding the error metrics throughout the whole year split up by land cover class as shown in table 1 RMSE/sd values stay below 1 for all forested classes and above 1 for arable land and pastures.

Table 1. Prediction errors that were externally validated against a testing data set from a different year, differentiated by land cover class (white rows). The gray row indicates the error from the cross validation within the model training.

Corine	Rsquared	RMSE	RMSE/sd
Mixed forest	0.29	1.25	0.84
Coniferous forest	0.20	1.21	0.90
Broad-leaved forest	0.25	1.44	0.88
Pastures	0.14	0.91	1.06
Arable land	0.05	0.75	1.46
All	0.40	1.12	0.78
All (cross validation error)	0.45	1.20	0.83

3.3. Spatio-temporal prediction

Spatial wall-to-wall predictions for the study area were computed based on monthly median values of Sentinel-1 and Sentinel-2 data from 2021. Figure 7 shows a sequence of four PAI predictions covering the growing season 2021. For a comparison of all 12 months see figure A1 in the appendix The general expectation that PAI increases significantly during the growing season and that there are significant differences between forest and non-forest are evident in figure 7. This impression is also confirmed by analyzing the predicted PAI dynamics separately for each land cover class (boxplots in figure 8, analogue to figure 3). Predictions during January

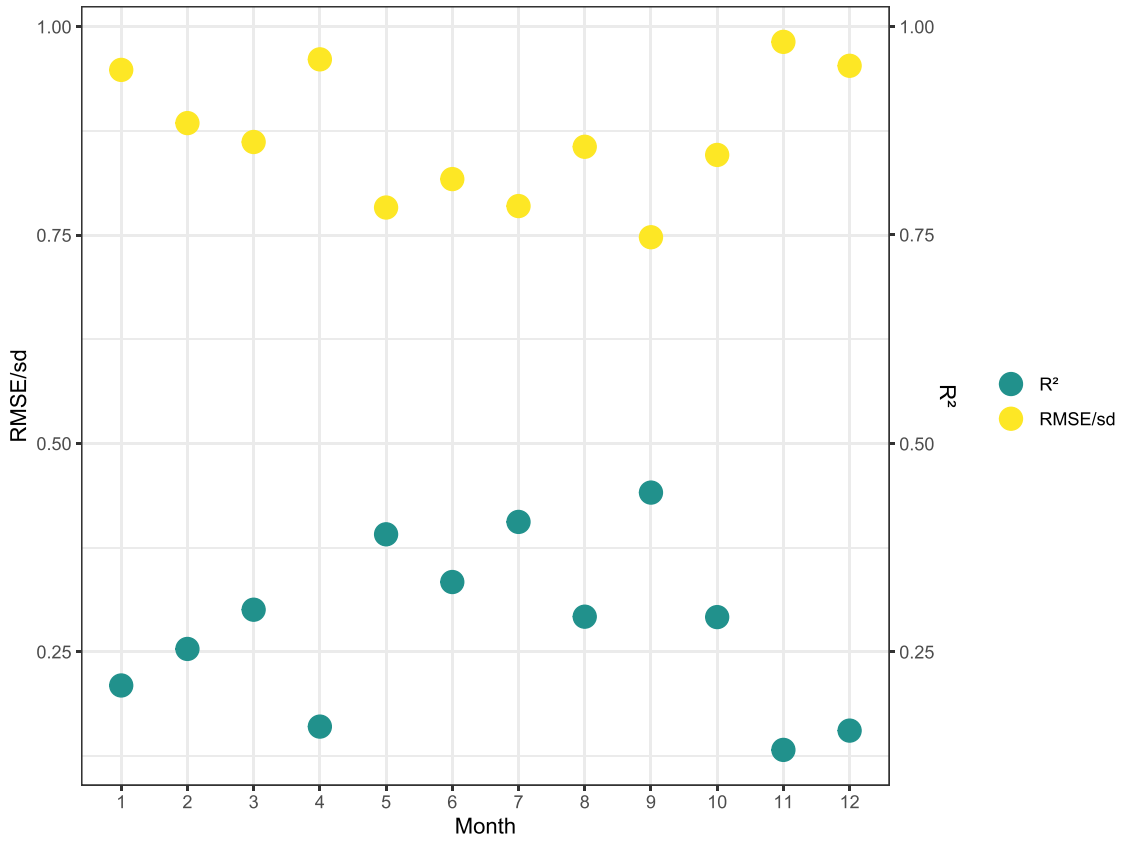


Figure 6. R2 and RMSE/sd error metrics for the 2021 validation data by month.

and November, however, are based on considerably less available pixels than during the rest of the year (figure 9). To check the agreement of the annual dynamics, Pearson's correlation coefficient was calculated between the monthly medians for the 2021 observations and the 2021 predictions. As detailed in table 2, this reveals a significant positive correlation for all forest classes but can not find a significant correlation for arable land and pastures.

4. Discussion

GEDI's PAI product generally reflects the expected temporal variability in the study area. In forested areas the increase in PAI during the early growing season corresponds to bud burst and leaf growth. On pastures, a slight gain together with an extended variability marks the start of the growing season. Seasonal patterns are more distinct for forested areas than for pastures and arable land. High variability of PAI on pastures during the summer months can be explained by irregular management activities such as mowing or grazing as well as isolated trees.

The spatial prediction of PAI also generally reflects the expected seasonal dynamics described in section 3.1 (see figures 7 and A1). PAI increases in forests (see figures 1 and 8) during spring and fluctuates in summer. The latter is probably the result of alternating extreme weather conditions in Germany. Related effects add to existing tree conditions induced by recent droughts (DWD 2019, 2020). For arable land and pastures, the temporal dynamic in figure 8 looks similar to the observed seasonal course (figure 3). For January and November the number of valid pixels is extremely low compared to the other months due to persistent and area-wide cloud coverage (see figures 9 and A1) and do therefore not allow for meaningful interpretation.

The PAI prediction was generally more accurate for the forest classes with broad-leaved forest and mixed forest performing even better than coniferous forest. This can likely be explained by the larger vertical variability of forests compared to other land cover types like arable land or pastures, and different moisture properties. Both GEDI LiDAR and the Sentinel sensors can capture differences in PAI more easily if vegetation height and structural variability are large. However, mixed forest shows inconsistency in performance throughout the year. This may either originate from its diverse ecological and structural composition and related spatio-temporal anomalies throughout the seasons, or from the relatively small number of reference footprints compared to the other land cover classes. Uncertainties in class assignment can be expected, especially from the delineation of pure deciduous or coniferous forest and mixed forest. This may be an additional factor causing lower model performance in the mixed forest class. In principle,

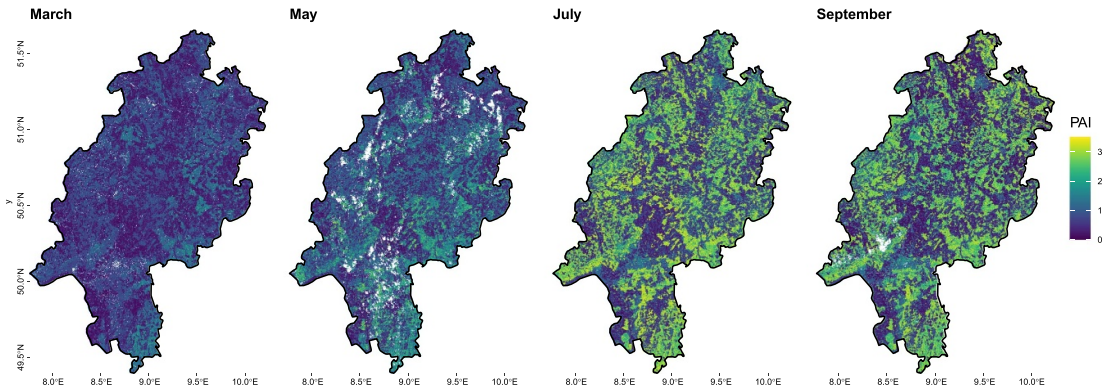


Figure 7. Spatial prediction of PAI for March, May, July and September for Hesse, Germany.

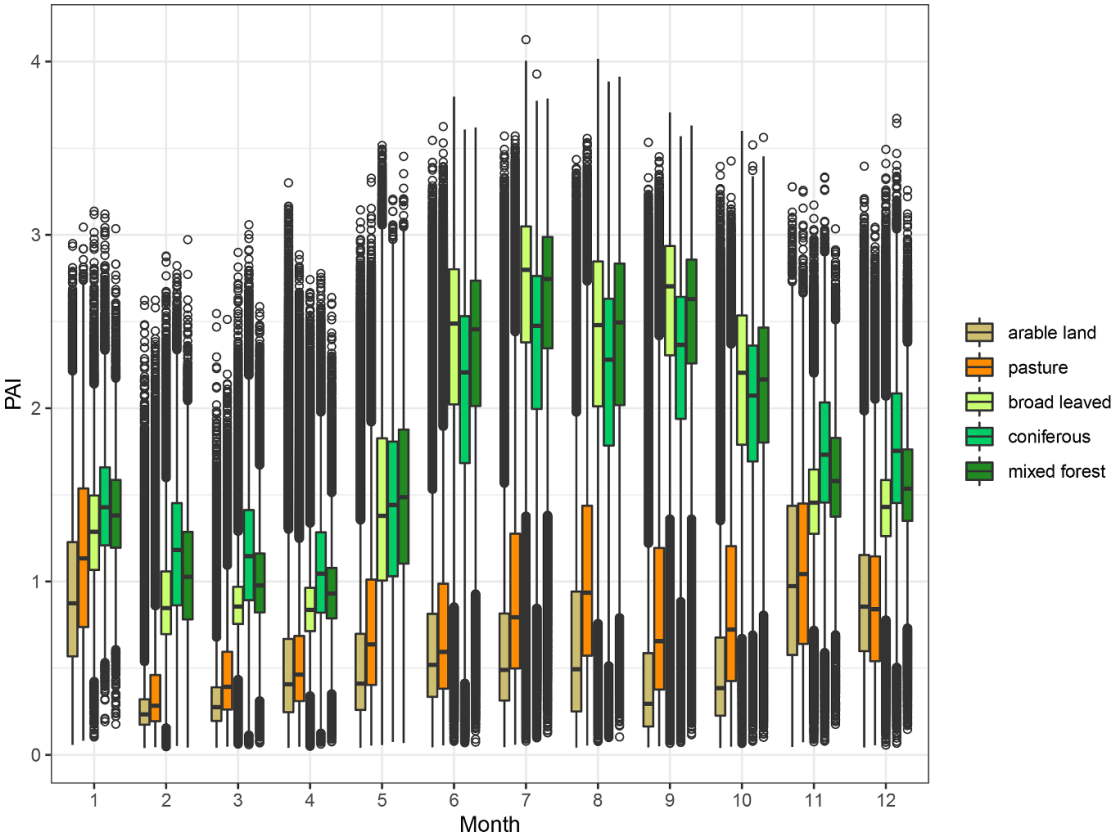


Figure 8. Temporal variability of the GEDI's PAI per land cover class including all pixels of monthly spatial predictions for 2021.

and despite the cross-check with the CORINE land cover classes, heterogeneous training footprints cannot be excluded from the analysis with absolute certainty. Since pastures and arable land in the study area are typically embedded in more heterogeneous structures (small settlements, roads, edges) that are not captured by CORINE, it is likely that these classes are more affected by impure footprints compared to larger homogeneous forest areas.

The temporal dynamics between observations and area-wide predicted pixel values for the test year 2021 agree well for forested areas. A significant positive correlation was found for the forested classes but not for arable land and pastures (see table 2). This may partly be explained by the overall smaller variance in PAI as well as individual growth and harvesting or mowing events of different crop types. Even though GEDI and Sentinel acquisition times can lie no further apart than three days, these areas can change significantly in the meantime, resulting in a mismatch between the data sets and, hence, an unfavorable effect on prediction quality. The monthly performance across all CORINE classes (see figure 6) shows a clear seasonal pattern with better performances during the summer months. The monthly RMSE/sd across all land cover classes scores below 1

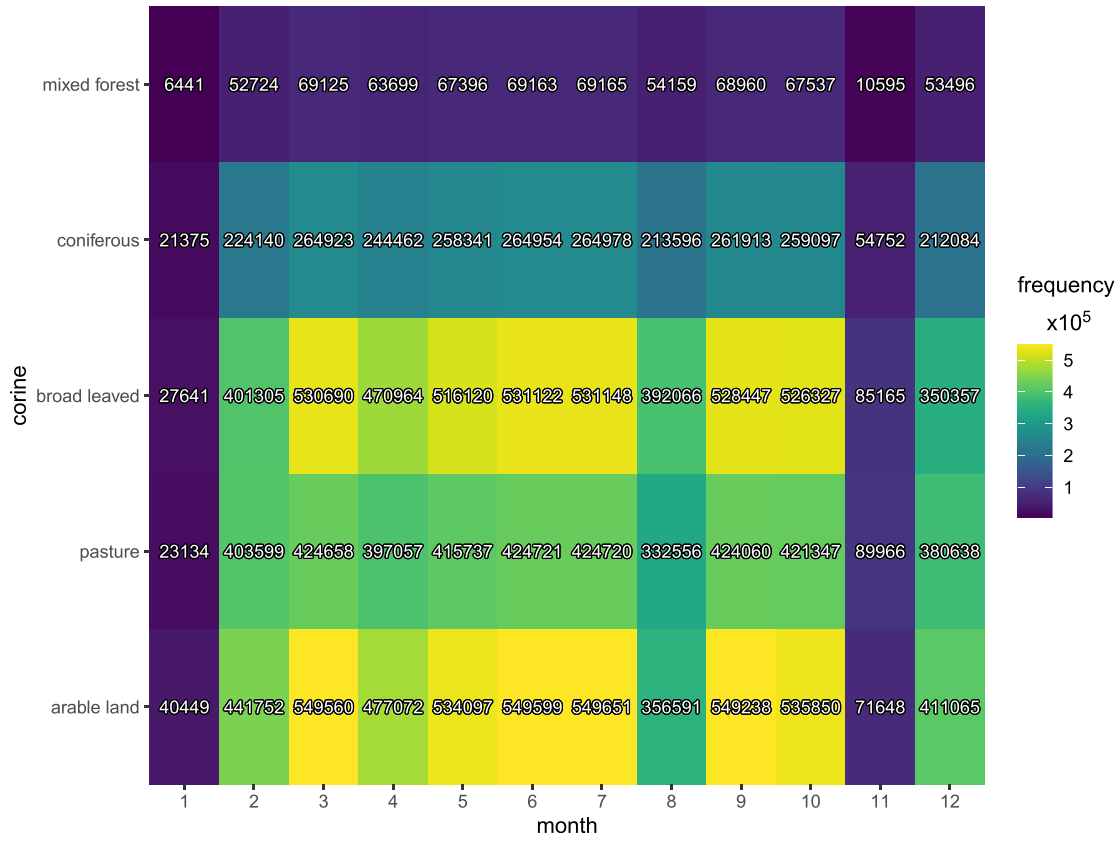


Figure 9. Number of available pixels in the monthly spatial predictions by land cover class.

Table 2. Pearson correlation of median monthly values between observations of testing data set and pixel values of area-wide spatial prediction. Significant results, in terms of the *p*-value, are marked bold.

Corine	Correlation	*p*-value
Arable land 12	0.018	0.956
Pastures 18	−0.004	0.990
Broad-leave forest 23	**0.910**	**0.000**
Coniferous forest 24	**0.729**	**0.007**
Mixed forest 25*	**0.821**	**0.002**

*Due to missing values in November for mixed forest, this value is only based on 11 median monthly values.

and therefore lower than the sd in the original data set. This proves that our model is generally usable throughout the year. January, April, November and December are the months with the weakest performance. This behavior may partially be attributed to sparse data availability and therefore extremely low RMSE/sd values for those months for arable land (see figure 5).

A performance comparison with other studies can only provide limited indications in this case. Most other studies used field-based measurements as reference and observed the related yet not identical LAI as the response variable. The accuracy of GEDI's PAI for different landscapes has only been explored in few studies. Dhargay *et al* (2022) tested the accuracy of GEDI's PAI in Australia and found rather poor agreement and a significant underestimation compared to estimates derived from air-borne LiDAR data. According to their assessment, however, both the complex study area and the time lag between air-borne and space-borne observations can be possible sources of error. They further used only about one month of GEDI data, which does not cover a time frame large enough to study temporal dynamics. Kacic *et al* (2021) examined the integration of GEDI's PAI and sentinel data in Paraguay's forests and aggregated the data across one entire dry season (RMSE = 0.3, R^2 around 0.5). Rishmawi *et al* (2021) produced contiguous PAI maps at 1 km resolution over the United States by integrating GEDI and VIIRS data (RMSE = 0.09, R^2 = 0.76). Since we use data with high spatial and temporal resolution, the performance of our models is worse, as would be expected, but difficult to relate directly. Few studies, including Miranda *et al* (2020), use PAI winter observations to calculate the woody proportion of forest areas. With this information they then calculate effective LAI during the growing season. This approach could potentially also be applied to GEDI studies and should be investigated further.

Other study design components, that complicate a direct comparison include, study areas, vegetation

types, study seasons, variance in the data, validation strategies and other aspects of the study design that vary considerably between publications. In addition, RMSE/sd values which facilitate a comparison across studies are rarely communicated and many previous studies restricted data analysis to plots that were clearly dominated by a single species (Brown *et al* 2019), or featured homogeneous cover (Korhonen *et al* 2017, Cohrs *et al* 2020). Nevertheless, it is helpful to put the results of this study in the context of previous research on LAI estimation. The GEDI models in our study generally reached lower R^2 values for arable land and pastures than models of previous studies. Frampton *et al* (2013) and Gitelson *et al* (2003) report a training error R^2 value range of 0.36 to 0.88 in their LAI models, differing drastically from the performance of our corresponding model (independently validated $R^2 = 0.05$). For pastures, Baghdadi *et al* (2016) presented R^2 values between 0.65 and 0.89 in contrast to 0.14 in our study. For agricultural areas, RMSE values of previous studies ranged from 0.44 to 0.68 (Delegido *et al* 2011, Verrelst *et al* 2015, Luo *et al* 2020) which is slightly better than our model (RMSE = 0.75). In the case of pastures we were able to obtain a lower error (RMSE = 0.91) compared to a study by Wang *et al* (2019) (RMSE = 1.09). For the forest classes, the GEDI-based approach achieved a slightly lower performance compared to other studies. While our RMSE scores are above 1.21, previous work by Korhonen *et al* (2017) and Meyer *et al* (2019b) reached values between 0.8 and 0.9 in models based on field-observations. The study of Cohrs *et al* (2020) in coniferous forest reached RMSE values of 0.63 to 0.89 with linear models while our validation reached an RMSE of 1.21. For deciduous forest, Brown *et al* (2019) achieved RMSE values of 1.55 using the Sentinel Application Platform (SNAP) algorithm, and up to 0.47 using an optimized algorithm, compared to an RMSE of 1.44 achieved in this study. In addition to previously described limitations in comparability, it is likely that the systematic, but spatio-temporally irregular coverage of GEDI footprints bares more challenges for the statistical modeling process compared to studies which do not use GEDI data. Luo *et al* (2020) for example used time series from a fixed set of plots and analyzed more homogeneous data than expected from the shifting GEDI footprints.

The different validation methods also have an impact on the comparability of the results. The spatial and temporal CV used in this study tests the applicability to new data, which in principle leads to poorer validation measures, but provides a realistic picture for prediction on new data. However, regardless of the absolute performance, the good agreement of error measures between the model-internal cross validation and the validation with the external test data set (see table 1) shows that modeling of LAI in regional studies is feasible for applications beyond the lifetime of the GEDI mission.

5. Conclusion

This study presents a first approach to match GEDI's PAI observations with their closest Sentinel-1 and -2 pixels in space and time to compile monthly wall-to-wall maps of PAI in heterogeneous landscapes. The high spatial resolution of the predictions and regular repetition rates improve the availability of information compared to current operational global LAI products. This study demonstrates that a stable year-round monitoring of the key vegetation variable PAI in a heterogeneous landscape is possible. However, our findings reveal that there are great differences in predictive power across land cover classes, the use of multi-temporal variables might be an option to optimize the model further. We further found the prediction of PAI within forests to be more stable possibly due to its variability in vertical structure. Overall our results show good agreement for predictions within the time range of our training data and one year beyond. This leads to the conclusion that our approach can even be applied to time periods outside GEDI's life-span.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/aliceziegler/GediEngineR.

Appendix

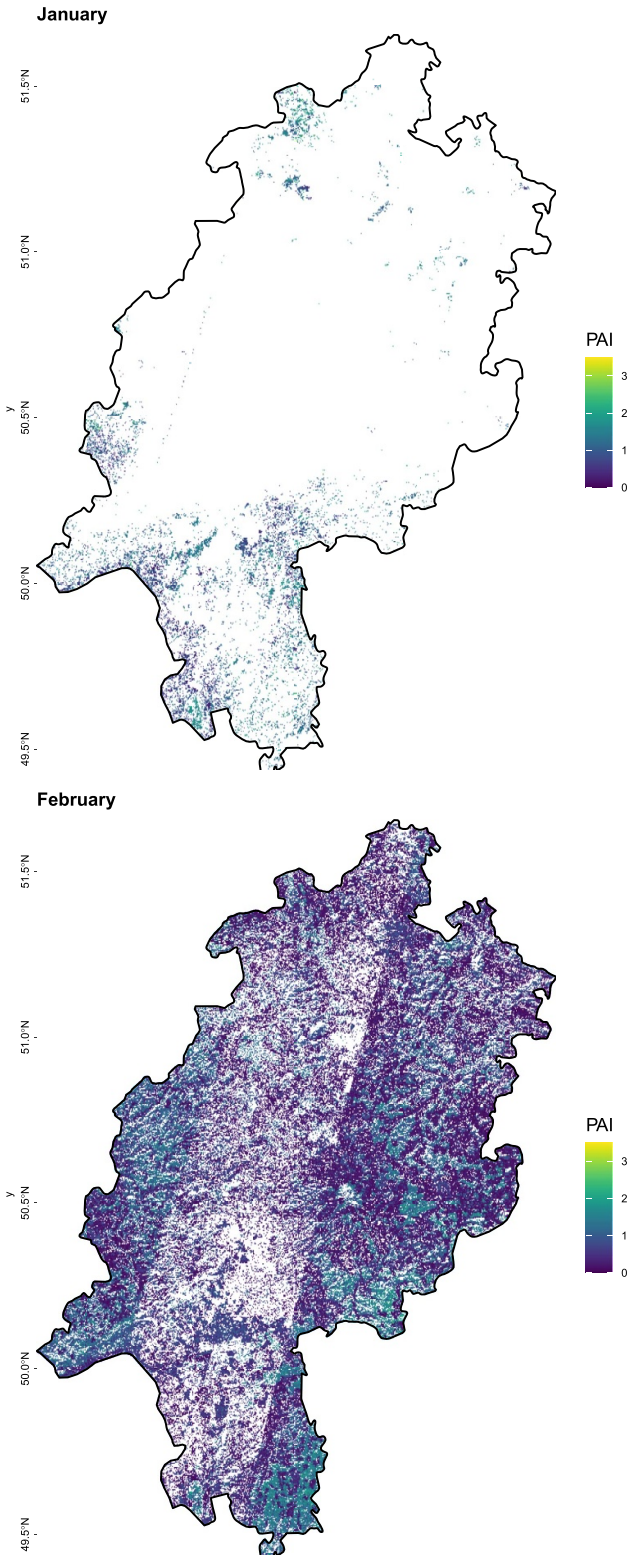


Figure A1. Monthly predictions of PAI in Hesse (Germany).

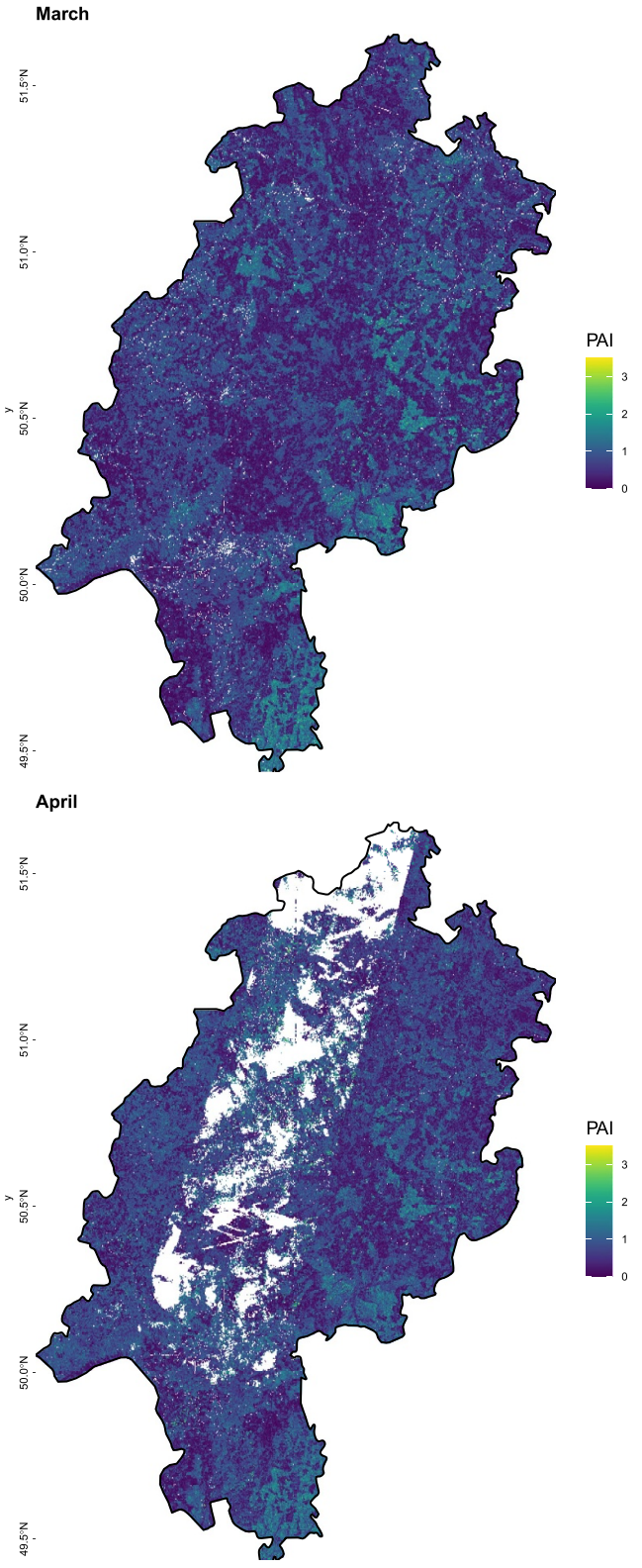


Figure A1. (Continued.)

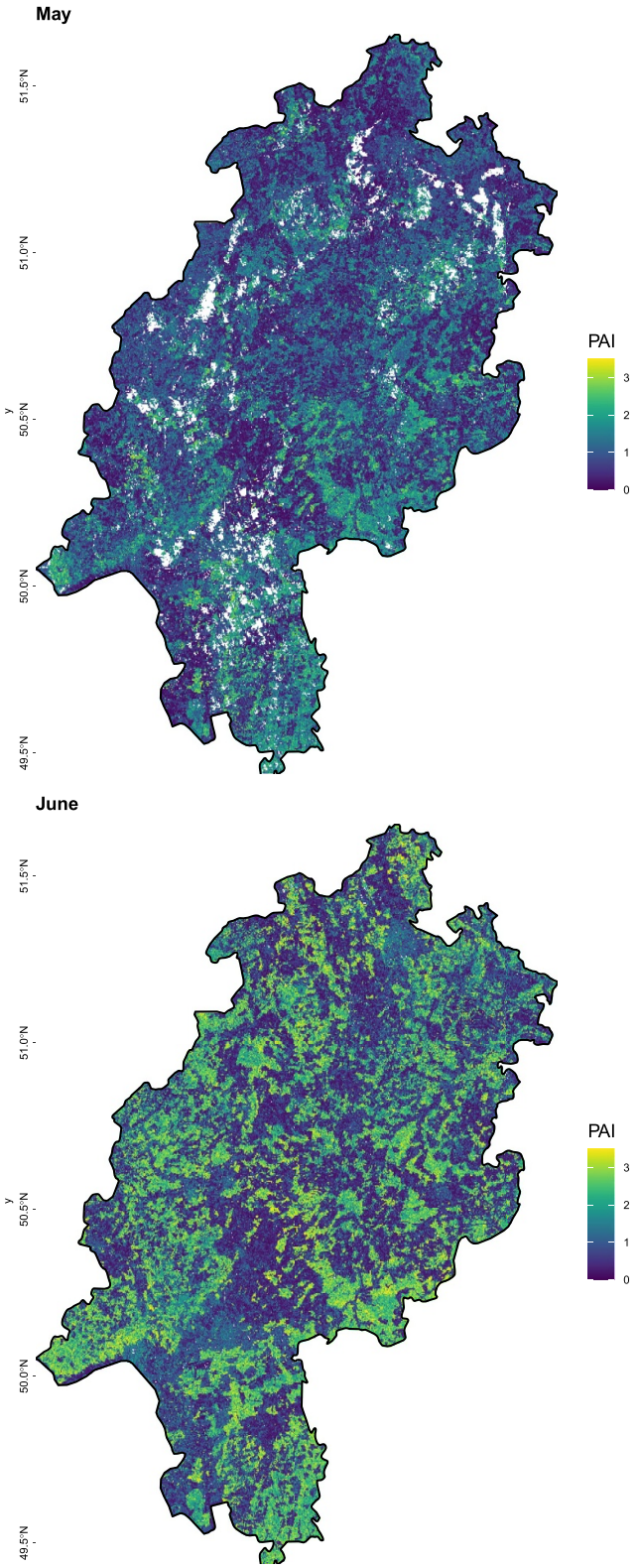


Figure A1. (Continued.)

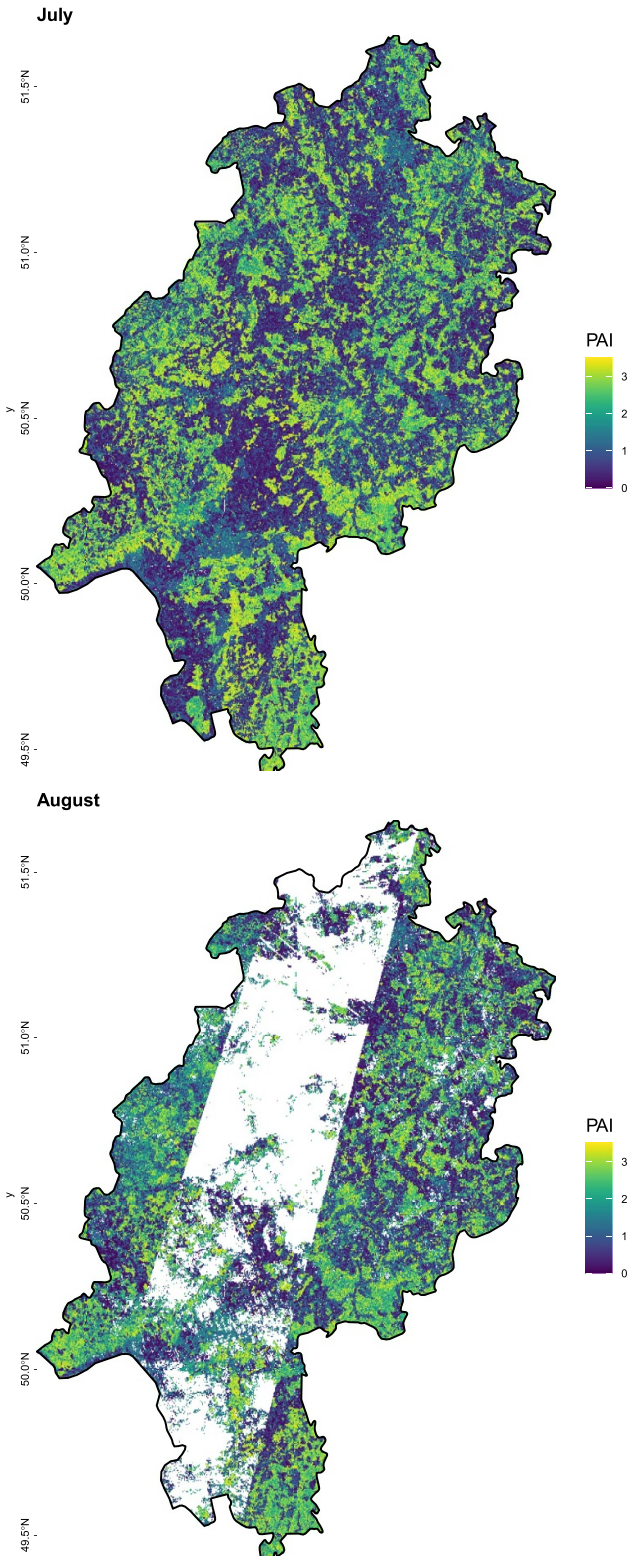


Figure A1. (Continued.)

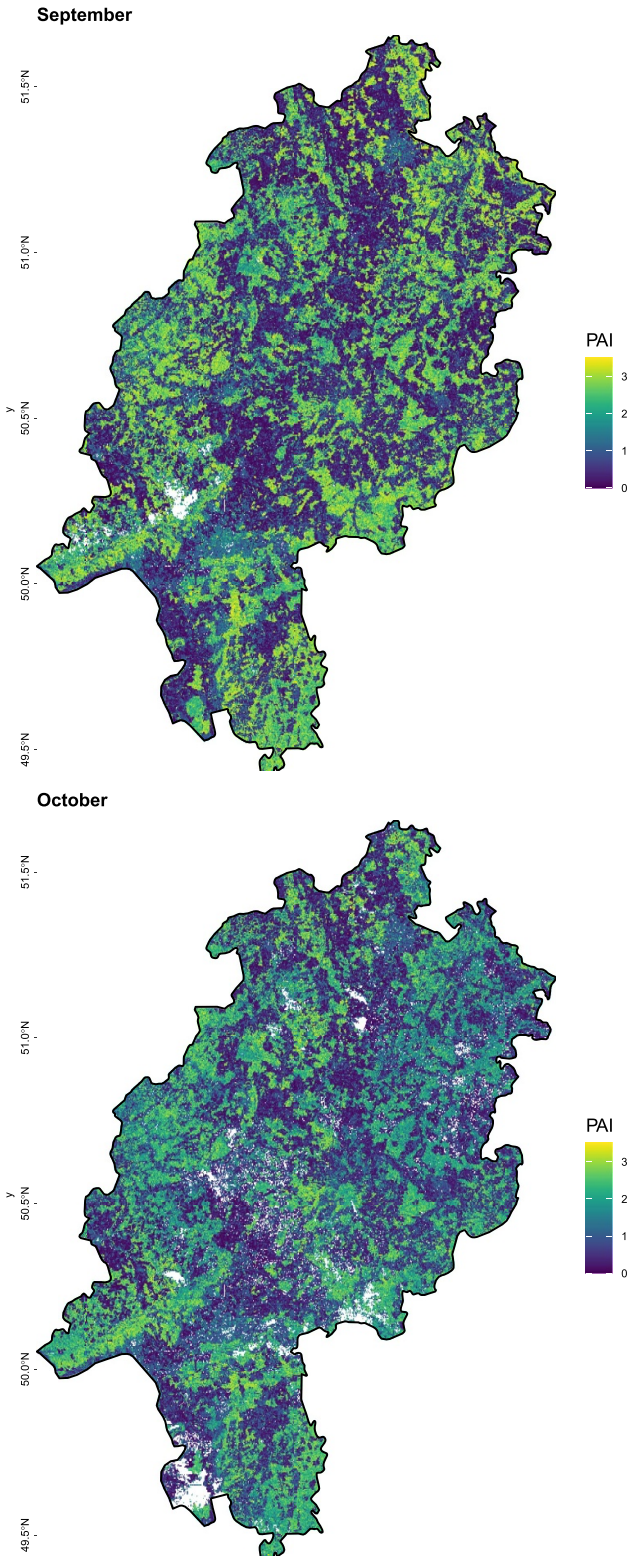


Figure A1. (Continued.)

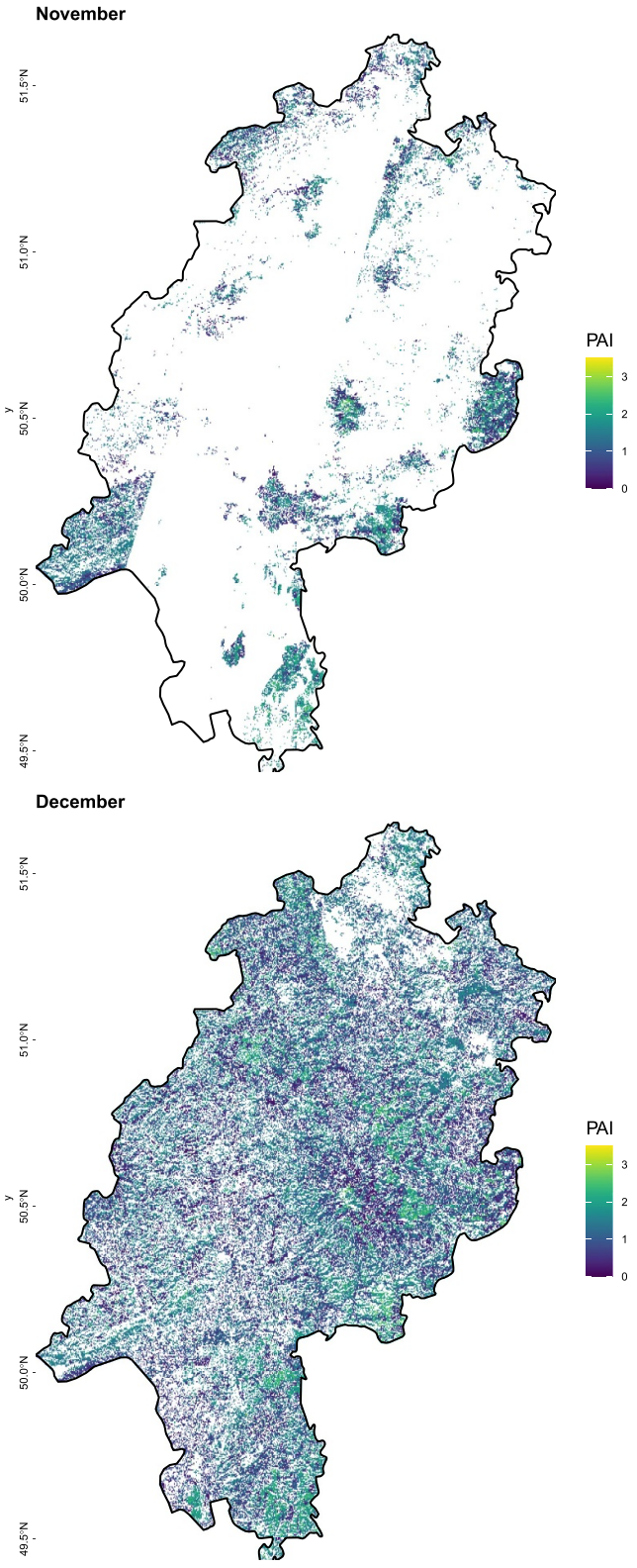


Figure A1. (Continued.)

ORCID iD

Alice Ziegler https://orcid.org/0000-0002-8613-8347

References

Bae S *et al* 2019 Radar vision in the mapping of forest biodiversity from space *Nat. Commun.* **10** 4757

Baghdadi N N, El Hajj M, Zribi M and Fayad I 2016 Coupling SAR C-band and optical data for soil moisture and leaf area index retrieval over irrigated grasslands *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **9** 1229–43

Boucher P B *et al* 2020 Detecting change in forest structure with simulated GEDI lidar waveforms: a case study of the Hemlock Woolly Adelgid (HWA; *Adelges tsugae*) infestation *Remote Sens.* **12** 1304

Brown L A, Ogutu B O and Dash J 2019 Estimating forest leaf area index and canopy chlorophyll content with Sentinel-2: an evaluation of two hybrid retrieval algorithms *Remote Sens.* **11** 1752

Chen L, Ren C, Zhang B, Wang Z, Liu M, Man W and Liu J 2021 Improved estimation of forest stand volume by the integration of GEDI LiDAR data and multi-sensor imagery in the Changbai Mountains Mixed forests Ecoregion (CMMFE), northeast China *Int. J. Appl. Earth Obs. Geoinf.* **100** 102326

Cohrs C W, Cook R L, Gray J M and Albaugh T J 2020 Sentinel-2 leaf area index estimation for pine plantations in the southeastern United States *Remote Sens.* **12** 1406

Delegido Jus, Verrelst J, Alonso L and Moreno J 2011 Evaluation of Sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content *Sensors* **11** 7063–81

Dhargay S, Lyell C S, Brown T P, Inbar A, Sheridan G J and Lane P N J 2022 Performance of gedi space-borne lidar for quantifying structural variation in the temperate forests of South-Eastern Australia *Remote Sens.* **14** 3615

di Tommaso S, Wang S and Lobell D B 2021 Combining GEDI and Sentinel-2 for wall-to-wall mapping of tall and short crops *Environ. Res. Lett.* **16** 125002

Dorado-Roda I, Pascual A, Godinho S, Silva C A, Botequim B, Rodriguez-Gonzalvez P, Gonzalez-Ferreiro E and Guerra-Hernandez J 2021 Assessing the accuracy of GEDI data for canopy height and aboveground biomass estimates in mediterranean forests *Remote Sens.* **13** 2279

Dubayah R *et al* 2020 The global ecosystem dynamics investigation: high-resolution laser ranging of the Earth's forests and topography *Sci. Remote Sens.* **1** 100002

DWD 2019 Jahrbuch 2019 (German Weather Service) (available at: www.dwd.de/DE/leistungen/jahresberichte_dwd/jahresberichte/2019.html;jsessionid=7FAE4634F4FC09C42A52F92FB92953AA.live21071?nn=511948)

DWD 2020 Jahrbuch 2020 (German Weather Service) (available at: www.dwd.de/DE/leistungen/jahresberichte_dwd/jahresberichte/2020.html;jsessionid=7FAE4634F4FC09C42A52F92FB92953AA.live21071?nn=511948)

Fang H, Baret F, Plummer S and Schaepman-Strub G 2019 An overview of global leaf area index (LAI): methods, products, validation and applications *Rev. Geophys.* **57** 739–99

Feret J-B, François C, Asner G P, Gitelson A A, Martin R E, Bidel L P R, Ustin S L, le Maire G and Jacquemoud Séphane 2008 PROSPECT-4 and 5: advances in the leaf optical properties model separating photosynthetic pigments *Remote Sens. Environ.* **112** 3030–43

Frampton W J, Dash J, Watmough G and James Milton E 2013 Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation *ISPRS J. Photogramm. Remote Sens.* **82** 83–92

Francini S, D'Amico G, Vangi E, Borghi C and Chirici G 2022 Integrating GEDI and landsat: spaceborne lidar and four decades of optical imagery for the analysis of forest disturbances and biomass changes in Italy *Sensors* **22** 2015

Frison P-L, Fruneau B, Kmiha S, Soudani K, Dufrêne E, Le Toan T, Koleck T, Villard L, Mougin E and Rudant J-P 2018 Potential of Sentinel-1 data for monitoring temperate mixed forest phenology *Remote Sens.* **10** 2049

Fuster B, Sánchez-Zapero J, Camacho F, García-Santos V, Verger A, Lacaze R, Weiss M, Baret F and Smets B 2020 Quality assessment of PROBA-V LAI, fAPAR and fCOVER collection 300 m products of copernicus global land service *Remote Sens.* **12** 1017

GCOS 2021 The status of the global climate observing system 2021: the GCOS status report (GCOS-240) *Status Report* (Geneva: GCOS)

Gitelson A A, Viña A, Arkebauer T J, Rundquist D C, Keydan G and Leavitt B 2003 Remote estimation of leaf area index and green leaf biomass in maize canopies *Geophys. Res. Lett.* **30** 1248

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D and Moore R 2017 Google Earth engine: planetary-scale geospatial analysis for everyone *Remote Sens. Environ.* **202** 18–27

Healey S P, Yang Z, Gorelick N and Ilyushchenko S 2020 Highly local model calibration with a new GEDI LiDAR asset on google Earth engine reduces landsat forest height signal saturation *Remote Sens.* **12** 2840

Huang B, Yang Y, Li R, Zheng H, Wang X, Wang X and Zhang Y 2022 Integrating remotely sensed leaf area index with biome-BGC to quantify the impact of land use/land cover change on water retention in Beijing *Remote Sens.* **14** 743

Jiang F, Smith A R, Kutia M, Wang G, Liu H and Sun H 2020 A modified KNN method for mapping the leaf area index in arid and semi-arid areas of China *Remote Sens.* **12** 1884

Kacic P, Hirner A and Da Ponte E 2021 Fusing Sentinel-1 and-2 to model GEDI-derived vegetation structure characteristics in GEE for the paraguayan chaco *Remote Sens.* **13** 5105

Kganyago M, Mhangara P, Alexandridis T, Laneve G, Ovakoglou G and Mashiyi N 2020 Validation of sentinel-2 leaf area index (LAI) product derived from SNAP toolbox and its comparison with global LAI products in an African semi-arid agricultural landscape *Remote Sens. Lett.* **11** 883–92

Khati U, Lavalle M and Singh G 2021 The role of time-series L-band SAR and GEDI in mapping sub-tropical above-ground biomass *Front. Earth Sci.* **9** 752254

Korhonen L, Hadi, Packalen P and Rautiainen M 2017 Comparison of Sentinel-2 and Landsat 8 in the estimation of boreal forest canopy cover and leaf area index *Remote Sens. Environ.* **195** 259–74

Kuhn M 2008 Building predictive models in R using the caret package *J. Stat. Softw.* **28** 1–26

Luo P, Liao J and Shen G 2020 Combining spectral and texture features for estimating leaf area index and biomass of maize using Sentinel-1/2 and Landsat-8 Data *IEEE Access* **8** 53614–26

Marselis S M *et al* 2020 Evaluating the potential of full-waveform lidar for mapping pan-tropical tree species richness *Glob. Ecol. Biogeogr.* **29** 1799–816

Meyer H 2020 CAST: 'caret' applications for spatial-temporal models (available at: https://github.com/HannaMeyer/CAST)

Meyer H, Reudenbach C, Hengl T, Katurji M and Nauss T 2018 Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation *Environ. Model. Softw.* **101** 1–9

Meyer H, Reudenbach C, Wöllauer S and Nauss T 2019a Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction *Ecol. Model.* **411** 108815

Meyer L H, Heurich M, Beudert B, Premier J and Pflugmacher D 2019b Comparison of Landsat-8 and Sentinel-2 data for

estimation of leaf area index in temperate forests *Remote Sens.* **11** 1160
Miranda R D Q, Nóbrega R L B, De Moura M S B, Raghavan S and Galvíncio J D 2020 Realistic and simplified models of plant and leaf area indices for a seasonally dry tropical forest *Int. J. Appl. Earth Obs. Geoinf.* **85** 101992
Myneni R B *et al* 2001 Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data *Remote Sens. Environ.* **83** 214–31
Padalia H, Sinha S K, Bhave V, Trivedi N K and Senthil Kumar A 2020 Estimating canopy LAI and chlorophyll of tropical forest plantation (North India) using Sentinel-2 data *Adv. Space Res.* **65** 458–69
Pasqualotto N, Delegido J, Van Wittenberghe S, Rinaldi M and Moreno J 2019 Multi-crop green LAI estimation with a new simple sentinel-2 LAI index (SeLI) *Sensors* **19** 904
Potapov P *et al* 2021 Mapping global forest canopy height through integration of GEDI and Landsat data *Remote Sens. Environ.* **253** 112165
Potapov P, Hansen M C, Kommareddy I, Kommareddy A, Turubanova S, Pickens A, Adusei B, Tyukavina A and Ying Q 2020 Landsat analysis ready data for global land cover and land cover change mapping *Remote Sens.* **12** 426
Qiao K, Zhu W, Xie Z and Li P 2019 Estimating the seasonal dynamics of the leaf area index using piecewise LAI-VI relationships based on phenophases *Remote Sens.* **11** 689
R Core Team 2022 *R: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing) (available at: www.R-project.org/)
Rishmawi K, Huang C, Schleeweis K and Zhan X 2022 Integration of VIIRS observations with GEDI-Lidar measurements to monitor forest structure dynamics from 2013 to 2020 across the conterminous United States *Remote Sens.* **14** 2320
Rishmawi K, Huang C and Zhan X 2021 Monitoring key forest structure attributes across the conterminous United States by integrating GEDI LiDAR measurements and VIIRS data *Remote Sens.* **13** 442
Seo H and Kim Y 2021 Role of remotely sensed leaf area index assimilation in eco-hydrologic processes in different ecosystems over east asia with community land model version 4.5 – biogeochemistry *J. Hydrol.* **594** 125957
Shendryk Y 2022 Fusing gedi with earth observation data for large area aboveground biomass mapping *Int. J. Appl. Earth Obs. Geoinf.* **115** 103108
Silva C A, Hamamura C, Valbuena R, Hancock S, Cardil A, Broadbent E N, de Almeida D R A, Silva Junior C H L and Klauberg C 2021 rGEDI: Nasa's Global Ecosystem Dynamics Investigation (Gedi) Data Visualization and Processing (available at: https://CRAN.R-project.org/package = rGEDI)
Tang H, Armston J, Hancock S, Marselis S, Goetz S and Dubayah R 2019 Characterizing global forest canopy cover distribution using spaceborne lidar *Remote Sens. Environ.* **231** 111262
Tang H, Dubayah R, Swatantran A, Hofton M, Sheldon S, Clark D B and Blair B 2012 Retrieval of vertical LAI profiles over tropical rain forests using waveform lidar at La Selva, Costa Rica *Remote Sens. Environ.* **124** 242–50
Tesemma Z K, Wei Y, Peel M C and Western A W 2015 The effect of year-to-year variability of leaf area index on variable infiltration capacity model performance and simulation of runoff *Adv. Water Resour.* **83** 310–22
Tharammal T, Bala G, Devaraju N and Nemani R 2019 A review of the major drivers of the terrestrial carbon uptake: model-based assessments, consensus and uncertainties *Environ. Res. Lett.* **14** 093005
Verhelst K, Gou Y, Herold M and Reiche J 2021 Improving forest baseline maps in tropical wetlands using GEDI-based forest height information and Sentinel-1 *Forests* **12** 1374
Verrelst J, Pablo Rivera J, Veroustraete F, Muñoz-Marí J, Clevers J G P W, Camps-Valls G and Moreno J 2015 Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods - A comparison *ISPRS J. Photogramm. Remote Sens.* **108** 260–72
Wang C, Elmore A J, Numata I, Cochrane M A, Lei S, Hakkenberg C R, Li Y, Zhao Y and Tian Y 2022 A framework for improving wall-to-wall canopy height mapping by integrating GEDI LIDAR *Remote Sens.* **14** 3618
Wang J, Xiao X, Bajgain R, Starks P, Steiner J, Doughty R B and Chang Q 2019 Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images *ISPRS J. Photogramm. Remote Sens.* **154** 189–201
Wang Y and Fang H 2020 Estimation of LAI with the LiDAR technology: a review *Remote Sens.* **12** 3457
Weiss M, Frederic B, Garrigues S and Lacaze R 2007 LAI and fAPAR CYCLOPES global products derived from VEGETATION. Part 2: validation and comparison with MODIS collection 4 products *Remote Sens. Environ.* **110** 317–31
Wright M N and Ziegler A 2017 A fast implementation of random forests for high dimensional data in C++ and R *J. Stat. Softw.* **77** 1–17 ranger
Xi Y, Tian Q, Zhang W, Zhang Z, Tong X, Brandt M and Fensholt R 2022 Quantifying understory vegetation density using multi-temporal sentinel-2 and gedi lidar data *GIsci. Remote Sens.* **59** 2068–83
Yan G, Hu R, Luo J, Weiss M, Jiang H, Mu X, Xie D and Zhang W 2019 Review of indirect optical measurements of leaf area index: recent advances, challenges and perspectives *J. Agric. Meteorol.* **265** 390–411
Zheng G and Moskal L M 2009 Retrieving leaf area index (LAI) using remote sensing: theories, methods and sensors *Sensors* **9** 2719–45
Ziegler A and Ludwig M 2023 Workflow to model GEDI PAI with Google Earth Engine and R *GEDIEngineR link* (available at: https://github.com/aliceziegler/GediEngineR)