Data in Brief 58 (2025) 111262

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

A comprehensive dataset of above-ground forest biomass from field observations, machine learning and topographically augmented allometric models over the Kashmir Himalaya

Syed Danish Rafiq Kashani, Faisal Zahoor Jan, Imtiyaz Ahmad Bhat, Nadeem Ahmad Najar, Irfan Rashid*

Department of Geoinformatics, University of Kashmir, Hazratbal Srinagar 190006, Jammu and Kashmir, India

ARTICLE INFO

Article history:
Received 9 December 2024
Revised 20 December 2024
Accepted 20 December 2024
Available online 25 December 2024

Dataset link: Data for: A comprehensive dataset of forest above-ground biomass from field observations and topographically augmented allometric models over the Kashmir Himalaya (Original data)

Keywords:
Forest carbon stock
Forest inventory
Remote sensing
Model hyperparameter optimization

ABSTRACT

Accurate estimates of forest dynamics and above-ground forest biomass for the topographically challenging Himalaya are crucial for understanding carbon storage potential, assessing ecosystem services, and guiding conservation efforts in response to climate change. This dataset provides a manually delineated multi-temporal forest inventory and a comprehensive record of above-ground biomass (AGB) across the Kashmir Himalaya, generated from field observations, advanced remote sensing and machine learning. Data were collected and generated through remote sensing techniques and extensive in-situ measurements of 6220 trees (n=275 plots), including tree diameter at breast height, species composition, and tree density to map forest area and model AGB across varied terrain. The dataset captures major forest types and species-specific AGB variation influenced by elevation, slope, and aspect. Additionally, newly developed species-specific allometric models, improved through the integration of normalized difference vegetation index (NDVI) and topographical augmentation are provided to improve AGB estimation accu-

* Corresponding author.
E-mail address: irfangis@kashmiruniversity.ac.in (I. Rashid).
Social media: @3poleresearcher (S.D.R. Kashani), @FaisalZahoorJa1 (F.Z. Jan), @bhatimtiyaz16 (I.A. Bhat), @A__nadeem (N.A. Najar), @irfansalroo (I. Rashid)

https://doi.org/10.1016/j.dib.2024.111262

racy. This dataset serves as a crucial resource for forest management, carbon monitoring, and ecological modeling, with broad applications in regional conservation strategies, biodiversity planning, and climate policy development in mountainous ecosystems.

Specifications Table

Subject	Agricultural Sciences
Specific subject areas	Forestry; Remote sensing; Ecological Modeling
Type of data	Image: GeoTIFF Vector: SHP Raw: XLSX Code: PDF Text: PDF
Data collection	Field data was collected between July and August of 2021. A total of 6220 individual trees were sampled across 275 plots, each measuring 20 × 20 meters (0.04 ha) in the Kashmir Himalaya (Fig. 1) following stratified random sampling approach covering coniferous, mixed and deciduous forest stands. Within these plots, information on tree characteristics such as diameter at breast height (dbh), species type, and density was documented. Sample locations were georeferenced using a Geomate SG7 GPS with a positional accuracy of ±3 m. Satellite data from Landsat (30 m) and Sentinel (10 m) available at https://earthexplorer.usgs.gov/ and https://www.sentinel-hub.com/explore/eobrowser/ respectively, were used to generate multi temporal forest inventory for the Kashmir Himalaya.
Data source location	Name of the study area: Kashmir Himalaya Districts: Anantnag, Bandipora, Baramulla, Budgam, Ganderbal, Kulgam, Kupwara, Pulwama, Shopain, and Srinagar Affiliation: University of Kashmir, Hazratbal Srinagar Jammu and Kashmir, India Country: India Latitude: 33°20′ N and 34°40′ N Longitude:73°40′ E and 75°40′ E
Data accessibility	Repository name: Zenodo Data identification number: 10.5281/zenodo.14329733 Direct URL to data: 10.5281/zenodo.14329733
Related research article	None

1. Value of the Data

- This baseline dataset is instrumental for researchers aiming to understand and characterize forest dynamics and AGB in the western Himalaya. This information supports climate change research by providing insights into the regional carbon dynamics and potential sequestration capacity of forests in the Kashmir Himalaya.
- The decadal-scale data on forest cover changes facilitate long-term ecological monitoring. Researchers can use this dataset to track changes in forest density and distribution over time, contributing to a better understanding of forest dynamics and ecological processes.
- The topographically-augmented and NDVI-based species-specific and general allometric models aid in addressing the challenges of AGB quantification in mountainous terrain, and can serve as a reference for studies in other high-altitude regions with similar ecological complexities (Fig. 1).
- The methodological workflow to mitigate inherent uncertainties in data and models developed in this dataset can be extended to other mountainous regions (Fig. 2). This allows for the initiation of long-term forest and biomass research in data-scarce areas, contributing to an accurate and broader understanding of global forest dynamics.

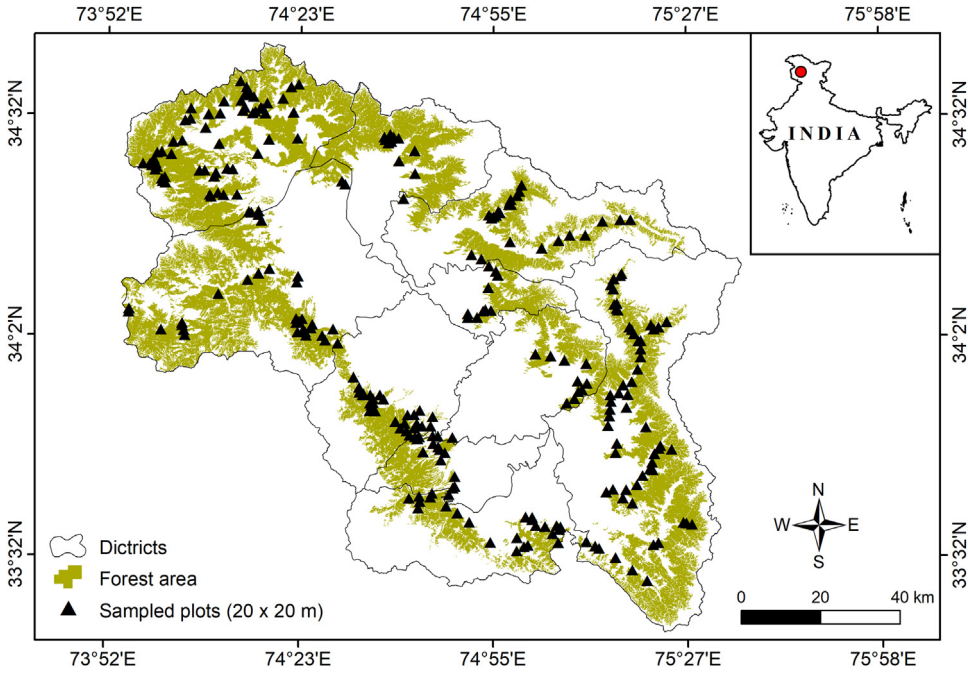

Fig. 1. Geographical distribution and location of the forest sampling plots across the study area.

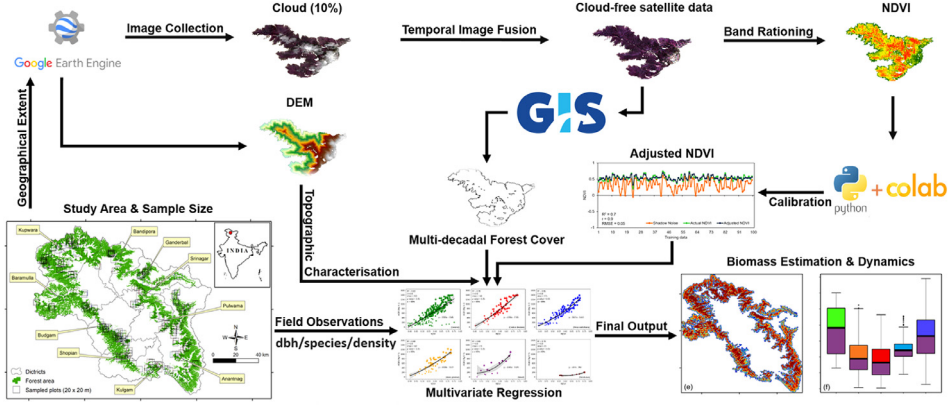

Fig. 2. A graphical representation of the comprehensive workflow adopted in this study.

2. Background

Forest ecosystems in mountainous regions like the Kashmir Himalaya play a vital role in supporting biodiversity, regulating water cycles, and sequestering carbon, making them crucial for both regional and global ecological health. These forests, however, face significant threats due to climate change, deforestation, and anthropogenic activities, leading to altered forest cover, reduced biomass, and compromised carbon storage potential. Despite their importance, forests in high-elevation terrains present considerable challenges for biomass estimation due to rugged topography, persistent cloud cover, and seasonal snow, limiting effective monitoring and conservation efforts. In response to these challenges, this dataset provides a comprehensive assessment

of forest dynamics and AGB across the Kashmir Himalaya, integrating data collected from extensive fieldwork, advanced remote sensing, and machine learning.

3. Data Description

Both raw data as well as processed data have been provided and uploaded to the Zenodo repository. The raw field-based individual tree-level measurements of 6220 trees are provided in a compressed folder as raw_data.rar. This raw data contains individual tree dbh, density, species, and AGB information of sample plots spread across 10 districts of Kashmir Himalaya. A separate Excel sheet for each district is provided as Anantnag.xlsx, Bandipora.xlsx, Baramulla.xlsx, Budgam.xlsx, Ganderbal.xlsx, Kulgam.xlsx, Kupwara.xlsx, Pulwama.xlsx, Shopain.xlsx, and Srinagar.xlsx. Furthermore, all equations used to calculate the individual tree level AGB are pre-inserted in the Excel sheets for public knowledge [8]. The processed data is provided as universally compatible GIS formats, including shapefiles and raster layers. In addition, the codes utilized in this study are provided as separate Python scripts in PDF format. The decadal forest cover inventories of Kashmir Himalaya for the years 1978, 1990, 2000, 2010, and 2021 are provided in shapefile format as FC1978.shp, FC1990.shp, FC2000.shp, FC2010.shp and FC2021.shp respectively. The forest inventory for the year 2021 provided as FC2021.shp also contains detailed forest cover classifications as attributes, allowing users to visualize and analyze density-wise forest distribution (degraded, dense, and sparse) across the study area. The district-level forest change estimates between 1978 and 2021 are provided in the shapefile format as District_FC_change.shp. Apart from this, the information on the study area and district geographical extents are also provided in shapefile format as Kashmir_himalaya_boundary.shp and District_boundaries.shp respectively. The scaled-up field-derived AGB data including the information on tree species and tree density for 275 sampled plots has been provided as Field_derived_AGB.shp. For reproducibility, the code to scale up the AGB estimates from a 20×20 m plot to a hectare is provided in a pdf format as Scaling_AGB.pdf. The parameters such as slope, aspect, and elevation utilized to understand the control of local topography on AGB are provided as slope.tif, aspect.tif, and elevation.tif respectively. Additionally, the decadal AGB raster layers generated with the help of a general allometric model developed as part of this study have been provided as AGB_1978.tif, AGB_1990.tif, AGB_2000.tif, AGB_2010.tif, and AGB_2021.tif. The code utilized to develop the general and species-specific models is also provided in a pdf format as regression_AGB.pdf. The loss in forest AGB that occurred due to forest degradation from 1978 to 2021 in the study area is provided in a raster format as AGB_loss.tif. An additional PDF file explaining the methodology adopted to generate and process the field observations and remotely sensed data is also provided as a supplementary document as methodology_adopted.pdf.

4. Experimental Design, Materials and Methods

This study employed a comprehensive approach to assess AGB across multiple decades in the Kashmir Himalaya. The dataset was generated through a combination of extensive field sampling, remote sensing analysis, and machine learning techniques to ensure high accuracy and applicability across varied forested landscapes (Fig. 2). High-resolution satellite data such as Sentinel 2A, Landsat 5 TM, Landsat 7 ETM+, Landsat 5 TM, and Landsat MSS was acquired for peak summer season (July to August) and pre-processed for the years 2021, 2010, 2000, 1990 and 1978 respectively at the Google Earth Engine platform. The images were filtered for the cloud cover with a 10% threshold with the help of the available metadata. Additionally, an advanced temporal image fusion technique was used to remove the inherent noise induced by the prevalent cloud cover in the Himalaya. For reproducibility, the GEE code for cloud cover fil-

tering and temporal image fusion is publicly accessible at https://code.earthengine.google.com/07b09571fbf550e096d747d4263a5b1a. The decadal forest cover inventories were generated utilizing the cloud-free satellite data and the on-screen digitization approach in a GIS environment at the scale of 1:10000 using the false-color composite images (NIR, Red, Green). The spatial resolution of the satellite data influences the accuracy of the manual digitization approach [1,2,3] therefore uncertainty in the area estimates was quantified as

$$E_A = n \times \lambda^2/2 \tag{1}$$

where n represents pixels along the perimeter, λ represents the spatial resolution of the utilized satellite data.

The density-based forest area classification for 2021 was conducted using both qualitative and quantitative methods, followed by statistical evaluation using field-based tree density data and Analysis of Variance (ANOVA) ($F = 598.29$, $p < 0.001$) [4]. A random sampling approach was conducted across the study area in the year 2021 from July to August to assess AGB at the plot level, with plot dimensions of 20 × 20 m. The locations of sampled plots were recorded using a Geomate SG7 device with a positional accuracy of ±3 m in handheld mode. 6220 trees were sampled encompassing 275 plots of different tree species across the Kashmir Himalaya. The species distribution across these plots was as follows: *Pinus walliciana* (120 plots), *Cedruis deodara* (75 plots), *Abies pindrow* (55 plots), Mixed coniferous (16 plots), and Decidous (9 plots). The circumference at breast height (cbh) was typically measured at 1.37 m above ground level [5–7] and later converted to the dbh using the standard mathematical relationship as:

$$dbh = \frac{cbh}{3.14} \tag{2}$$

Trees with cbh <31.5 cm were not sampled and considered as scrubs. Since the topographic and climatic conditions of the Gilgit-Baltistan and Kashmir Himalaya are comparable, this study used the allometric models developed in the Gilgit-Baltistan through the destructive method to estimate the individual and plot level AGB in the study area [8]. The AGB estimates are associated with a gross error of 11.36 %, incorporating measurement, sampling, and allometric model errors. This estimation follows the additive method proposed and implemented by [9,10]. Field-based tree density and AGB information at the plot level were used to quantify the AGB per hectare. The information on the region topography including slope, aspect, and elevation was generated from the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) with a spatial resolution of 90 m to understand the topographic influence and reveal patterns in the AGB. In addition, the NDVI was generated from the Sentinel and Landsat data using the following formula:

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{3}$$

where *NIR* and *RED* represent reflectance values in NIR and RED bands of satellite data respectively.

To address residual noise resulting from shadow cover in the NDVI, the calibration was performed using the Random Forest machine learning algorithm. To train the model, the data was split in the ratio of 80: 20. A hyperparameter tuning was performed to get the best possible settings and improve the model prediction. The model performed better with a correlation coefficient of 0.9, a coefficient of determination of 0.7, and a root mean square error of 0.05 at the following settings: random state = 10, number of estimators = 200, test size = 0.20, and maximum depth = 20. The trained model (rf.model.joblib) has been made freely available at the Zenodo repository for public reproducibility. Field-based AGB measurements were then correlated with the calibrated NDVI and local topographic factors, including slope, aspect, and elevation, to develop improved species-specific regression-based allometric models for the study area. The AGB for 2021, 2010, 2000, 1990, and 1978 was estimated using the developed general allometric model utilizing topographic variables and NDVI information having a correlation coefficient and coefficient of determination of 0.8 and 0.63 respectively. The general and species-specific allometric models developed in this study are given in Table 1.

Table 1
General and species-specific allometric models developed using the field-observed AGB and remote sensing datasets.

Model/Species	Statistical parameters for the model										
	α_1	α_2	α_3	α_4	c	Sample size (number)	dbh range (cm)	Density range (number/0.04ha)	r	r^2	*RMSE*
General	−0.897	−0.01	−0.06	3461	−1240	275	14–153	3–153	0.80	0.63	± 83.9
Cedrus deodara	−0.307	−0.02	−0.25	3391	−1139	75	14–77	5–72	0.88	0.78	± 145
Pinus wallichiana	0.91	0.03	0.07	3497	−1550	120	17–92	9–153	0.92	0.86	± 122.8
Abies pindrow	2.07	−0.14	0.21	2799	−1598	55	19–153	3–56	0.85	0.72	± 151.4
Mixed coniferous	4.93	−0.13	0.39	2663	−1880	16	20–86	13–68	0.89	0.80	± 133.7
Deciduous	−0.25	0.24	−0.12	232	186	9	18–47	17–42	0.93	0.88	± 16.5

Equations are in the form of $y = \alpha_1 \times \beta_1 + \alpha_2 \times \beta_2 + \alpha_3 \times \beta_3 + \alpha_4 \times \beta_4 + c$. Where, α and c are the slope and y-intercept of the regression model. β_1, β_2, β_3, and β_1 are the slope, aspect, elevation, and normalized difference vegetation index information respectively. The r and r^2 and *RMSE* are the Pearson correlation, coefficient of determination, and root mean square error calculated through multivariate statistical analysis respectively.

Limitations

The rugged terrain and frequent cloud cover in the Himalaya restrict data collection during certain periods, potentially limiting the temporal continuity of remote sensing imagery. Although this study addresses these challenges by employing advanced remote sensing approaches such as temporal image fusion, and machine learning, some small-scale forest patches might not be fully represented in high-elevation shadowed areas. Additionally, enhanced species-specific allometric models developed in this study lack field measurements from inaccessible regions, which may not fully capture the ecological variability across remote and less accessible areas. The unique microclimates and diverse forest types may lead to variability in the AGB estimates which could be further refined by expanding the sampling density in future research. Despite these limitations, the dataset provides valuable baseline information and robust AGB estimates, though caution is advised in interpreting results for small or highly localized areas. Future work can focus on enhancing model precision with additional field data and incorporating more advanced spectral indices to address these limitations further.

Ethics Statement

No human or animal studies are presented in the manuscript. The study does not involve any data that was acquired from any social media platforms.

CRediT Author Statement

Syed Danish Rafiq Kashani: Original draft preparation, Data curation, Visualization, Investigation, Writing - Reviewing and Editing. **Faisal Zahoor Jan:** Visualization, Investigation, Writing - Reviewing and Editing. **Imtiyaz Ahmad Bhat:** Visualization, Investigation, Writing - Reviewing and Editing. **Nadeem Ahmad Najar**: Visualization, Investigation, Writing - Reviewing and Editing. **Irfan Rashid**: Original draft preparation, Supervision, Conceptualization, Writing - Reviewing and Editing.

Data Availability

Data for: A comprehensive dataset of forest above-ground biomass from field observations and topographically augmented allometric models over the Kashmir Himalaya (Original data) (Zenodo)

Acknowledgments

The authors express gratitude to USGS and ESA for freely hosting Landsat and Sentinel data. The authors also acknowledge the use of Google Colab and Google Earth Engine for the processing and analysis. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Bolch, B. Menounos, R. Wheate, Landsat-based inventory of glaciers in western Canada, 1985–2005, Remote Sens. Environ. 114 (2010) 127–137, doi:10.1016/j.rse.2009.08.015.
[2] D.K. Hall, K.J. Bayr, W. Schöner, R.A. Bindschadler, J.Y. Chien, Consideration of the errors inherent in mapping historical glacier positions in Austria from the ground and space (1893–2001), Remote Sens. Environ. 86 (2003) 566–577, doi:10.1016/S0034-4257(03)00134-2.
[3] F. Paul, N.E. Barrand, S. Baumann, E. Berthier, T. Bolch, K. Casey, H. Frey, S.P. Joshi, V. Konovalov, R. Le Bris, On the accuracy of glacier outlines derived from remote-sensing data, Ann. Glaciol. 54 (2013) 171–182, doi:10.3189/2013AoG63A296.
[4] S.F. Sawyer, Analysis of variance: the fundamental concepts, J. Manual Manipul. Therapy 17 (2009) 27E-38E., doi:10.1179/jmt.2009.17.2.27E.
[5] J.D. Waskiewicz, L.S. Kenefic, N.S. Rogers, J.J. Puhlick, J.C. Brissette, R.J. Dionne, in: Sampling and Measurement Protocols for Long-term Silvicultural Studies on the Penobscot Experimental Forest, US Department of agriculture, Forest Service, Northern Research Station, Newtown Square, PA, 2015, p. 32, doi:10.2737/NRS-GTR-147. General Technical Report NRS-147147, 1–32.
[6] D. Wang, B. Wan, P. Qiu, Z. Zuo, R. Wang, X. Wu, Mapping height and aboveground biomass of mangrove forests on Hainan Island using UAV-LiDAR sampling, Remote Sens. 11 (18) (2019) 2156, doi:10.3390/rs11182156.
[7] N.C. Swayze, W.T. Tinkham, Application of unmanned aerial system structure from motion point cloud detected tree heights and stem diameters to model missing stem diameters, MethodsX 9 (2022) 101729, doi:10.1016/j.mex.2022.101729.
[8] A. Ali, Biomass and Carbon Tables for Major Tree Species of Gilgit Baltistan, Pakistan Forest Institute Peshawar, Pakistan, 2017 https://fwegb.gov.pk/wp-content/uploads/2022/01/BIOMASS-TABLES-OF-CONIFEROUS-SPECIES-OF-GILGIT-BALTISTAN-Revised-1-1.pdf.
[9] J. Chave, R. Condit, S. Lao, J.P. Caspersen, R.B. Foster, S.P. Hubbell, Spatial and temporal variation of biomass in a tropical forest: results from a large census plot in Panama, J. Ecol. 91 (2003) 240–252, doi:10.1046/j.1365-2745.2003.00757.x.
[10] M. Urbazaev, C. Thiel, F. Cremer, R. Dubayah, M. Migliavacca, M. Reichstein, C. Schmullius, Estimation of forest aboveground biomass and uncertainties by integration of field measurements, airborne LiDAR, and SAR and optical satellite data in Mexico, Carbon Bal. Manag. 13 (2018) 5, doi:10.1186/s13021-018-0093-5.