scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A Global Dataset of Location Data Integrity-Assessed Reforestation Efforts

Angela John[1]✉, Selvyn Allotey[1], Till Koebe[1], Alexandra Tyukavina[2] & Ingmar Weber[1]

Afforestation and reforestation are popular strategies for mitigating climate change by enhancing carbon sequestration. However, the effectiveness of these efforts is often self-reported by project developers, or certified through processes with limited external validation. This leads to concerns about data reliability and project integrity. In response to increasing scrutiny of voluntary carbon markets, this study presents a dataset on global afforestation and reforestation efforts compiled from primary (meta-)information and augmented with time-series satellite imagery and other secondary data. Our dataset covers 1,289,068 planting sites from 45,628 projects spanning 33 years. Since any remote sensing-based validation effort relies on the integrity of a planting site's geographic boundary, this dataset introduces a standardized assessment of the provided site-level location information, which we summarize in one easy-to-communicate key indicator: LDIS – the Location Data Integrity Score. We find that approximately 79% of the georeferenced planting sites monitored fail on at least 1 out of 10 LDIS indicators, while 15% of the monitored projects lack machine-readable georeferenced data in the first place. In addition to enhancing accountability in the voluntary carbon market, the presented dataset also holds value as training data for e.g. computer vision-related tasks with millions of linked Sentinel-2 satellite images.

Background & Summary

With forests being one of the major carbon sinks on our planet, there have been numerous efforts to protect, restore and extend forested areas to mitigate climate change[1,2].However, while economic rent seeking through depletion of natural resources such as unsustainable logging, overgrazing, conversion to farmland,establishment of monoculture plantations, and mismanaged firescan limit the effectiveness of these efforts, forests also offer renewable resources (e.g., sustainably managed timber), highlighting the importance of management practices that balance resource use with longterm ecosystem integrity[3,4]. The rise of the voluntary carbon market (VCM) has provided landowners with an alternative source of income that aligns restoration goals with economic incentives. In contrast to heavily regulated compliance carbon markets such as the European Union Emissions Trading System, VCM prices per ton of CO_2 offset are mainly determined by market forces of supply and demand. Public controversies that have sparked in recent years around the additionality (see box 'Key Terms') of VCM activities, especially of REDD+ projects, have depressed buyer confidence, reducing demand and market volumes in parts of the VCM since 2021. Moreover, recent analyses[5,6] demonstrate that many REDD+ projects overstate additionality and fail to deliver the claimed emission reductions, further undermining buyer confidence and contributing to price declines[7–12].

[1]Saarland Informatics Campus, Department of Computer Science, Saarbrücken, 66123, Germany. [2]University of Maryland, Department of Geographical Sciences, Riverdale, 20737, USA. ✉e-mail: ajohn@cs.uni-saarland.de

Key Terms

1. **Article 6 of the Paris Agreement:** The Paris Agreement, adopted by the Parties to the United Nations Framework Convention on Climate Change (UNFCCC), aims to limit global temperature rise to below 2°C above pre-industrial levels, with efforts to keep it to 1.5°C. It emphasizes equity and common but differentiated responsibilities to support sustainable development and poverty eradication. It also promotes the conservation of carbon sinks, such as forests, and encourages actions to reduce emissions from deforestation. Countries must submit their nationally determined contributions (NDCs) every five years, subject to expert review. The Agreement came into force on November 4, 2016, and by April 2021, it had been adopted by 194 Parties.
2. **REDD+:** A UNFCCC-developed voluntary climate change mitigation framework aimed to facilitate developing countries in "Reducing Emissions from Deforestation and forest Degradation", with "+" emphasizing the role of conservation, sustainable management of forests, and enhancement of forest carbon stocks. REDD+ can be implemented at multiple levels, from specific project areas (e.g., a forest concession or protected area) to broader subnational or national initiatives.
3. **Voluntary carbon markets:** A voluntary carbon market (VCM) is a decentralised platform where private entities can buy and sell carbon credits that represent removals or reductions of greenhouse gases (GHGs) in the atmosphere in avoluntarily beyond regulatory compliance frameworks.
4. **Permanence:** In the context of carbon projects, permanence refers to the durability of stored carbon while acknowledging the risk of reversal; it is not a binary attribute but a continuum managed over time through measures such as long-term monitoring, insurance, and other risk mitigation strategies to maintain carbon benefits over extended periods.
5. **Additionality:** The requirement that the emissions reductions or carbon sequestration achieved by a REDD+ or carbon offset project are greater than what would have occurred in the absence of the project.
6. **Verification:** The independent assessment and confirmation of the emissions reductions or carbon sequestration claimed by a carbon offset project, carried out by third-party auditors who evaluate whether the project's reported outcomes are accurate, credible, and consistent with the standards set by a given carbon market protocol.
7. **Deforestation:** Deforestation describes the process of clearing forest from an area and converting the area to non forest(agriculture,bare land,infrastructure etc.).
8. **Reforestation:** Reforestation describes the process of replanting trees in areas where forests have decreased previously (within the past 50-100 years) due to cleared or degraded ecosystems usually through deforestation.
9. **Afforestation:** The process of planting trees or sowing seeds of trees in areas where previously no trees existed within the past 50-100 years.When applied to degraded or deforested lands, afforestation can provide carbon sequestration and ecosystem benefits; however, planting trees in naturally non-forested ecosystems (e.g., native grasslands, savannas, peatlands) can cause substantial biodiversity loss and alter ecosystem functions. Thus, site selection and ecosystem context are critical when considering afforestation.
10. **Forest:** According to Food and Agriculture Organization of the United Nations,[12] "a forest is an area of land of more than 0.05 hectares with a tree canopy cover of at least 10% with a minimum height between two and five meters. In our work, we use GLAD forest with the forest definition extended to include the trees outside the forest, such as those in agroforestry areas". We acknowledge, however, that what constitutes a "forest" is the subject of ongoing debate. Multiple, sometimes conflicting, definitions exist in the literature.[1]

A major limitation of current reforestation initiatives is the lack of independent, high-quality spatial data to verify a project's location data integrity. Here, by 'integrity' we mean the validity, accuracy and completeness of the spatial information—ensuring that the site's mapped boundaries are topologically valid (no self-intersections), free of spurious overlaps or gaps, and closely correspond to the true extent of the on-the-ground reforestation efforts. Many datasets rely on self-reported information from project developers or certification bodies, which often lacks external validation. Certification frameworks such as Verra (https://registry.verra.org/), Gold Standard (https://www.goldstandard.org/),and the American Carbon Registry (https://wppremiumplugins.com/americancarbonregistry/) offer structured methodologies but substantially vary in rigor, leading to inconsistencies in credit quality[13]. These inconsistencies complicate the comparison and aggregation of carbon credits, making it difficult for buyers to assess and ensure the integrity of their offsets. As

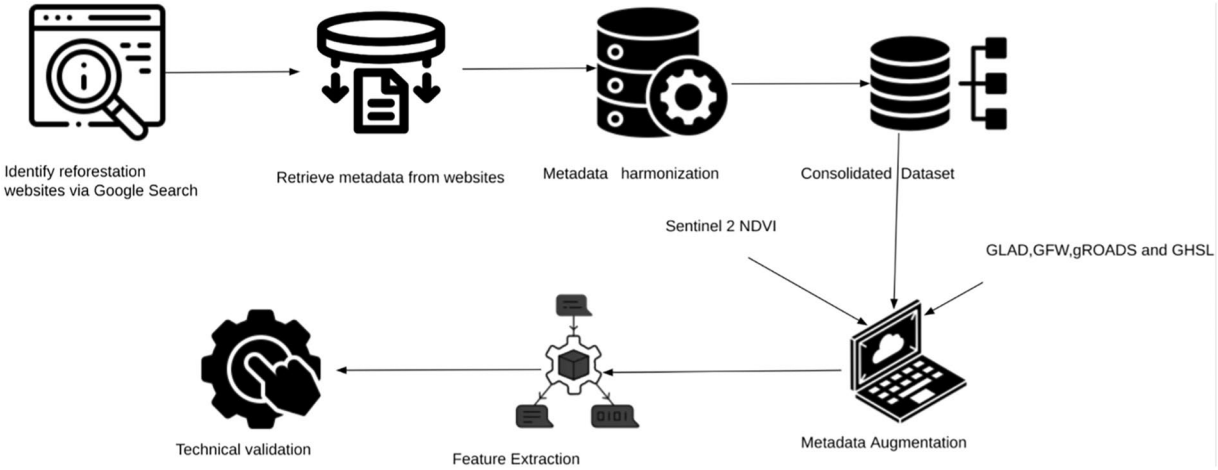

Fig. 1 The reforestation data generation workflow. We follow these steps in our reforestation data collection.

Index	Definition/Description	Formula
NDVI (Normalized Difference Vegetation Index)	Provides information about the presence or absence of green vegetation, and can indicate vegetation health in some cases.	$\frac{\text{Near Infrared (NIR)} - \text{Red}}{\text{Near Infrared (NIR)} + \text{Red}}$
NDRE	Identifying subtle stress factors and chlorophyll variations in trees	$\frac{\text{Near Infrared (NIR)} - \text{RedEdge}}{\text{Near Infrared (NIR)} + \text{RedEdge}}$
SAVI	To correct NDVI for the influence of soil brightness in areas where vegetative cover is low.	$\frac{\text{NIR} - \text{RedEdge}}{\text{NIR} + \text{RedEdge} + 0.5} * (1+0.5)$

Table 1. Overview of NDVI, NDRE and SAVI definitions. Using bands from Sentinel-2.

the implementation of Article 6 of the Paris Agreement (see box 'Key Terms') advances, ensuring transparency and standardization in VCM becomes increasingly critical. The major carbon crediting companies have recently responded by joining global compliance efforts such as the United Nations' International Civil Aviation Organization (ICAO) CORSIA[14] or the Core Carbon Principles (CCPs)[15], but there is still a long way to go[10].

To address these challenges, this study introduces a global dataset of quality-assessed, georeferenced reforestation projects. The dataset consolidates information from over 50 sources, covering 1.29 million planting sites from 45,628 projects spanning 33 years. In addition to the core location and project metadata, we augment information about each reforestation site with periodic Sentinel-2 satellite imagery and with secondary datasets capturing infrastructure presence, land-cover transitions, and local climatic conditions. We also use large-language models to extract contextual details directly from project descriptions and accompanying project documentation. Finally, we derive integrity indicators from these variables to flag potential issues with the reported site locations.

The list of websites from which the data are acquired can be found in Table 7 in the Appendix. An overview of the indicators available in the dataset can be found in Table 2 and in Tables 8 and 9 in the Appendix. Figure 1 describes the underlying data aggregation process.

Assessing the success of reforestation efforts is inherently multidimensional, encompassing ecological, social, and economic factors[16–19]. Various indicators have been proposed to evaluate different aspects of success: establishment success (e.g. seedling survival rates, initial tree growth), forest growth success (e.g., stand density, biomass accumulation), environmental success (e.g., biodiversity restoration, soil stabilization, carbon sequestration) and socioeconomic success (e.g., employment generation, community engagement, land tenure security). However, these indicators are rarely standardized across projects and many rely on self-reported data with limited independent verification[20]. Remote sensing offers a scalable approach to monitoring tree cover change[21–28], yet it struggles to capture more nuanced success metrics such as species composition, ecosystem resilience, or socioeconomic benefits[29–31].

However, every remote sensing-based monitoring effort depends on the accuracy and validity of the geographical boundaries of the monitored planting sites. Our location data integrity assessment considers multiple spatial indicators to identify potential discrepancies or inaccuracies in the georeferenced data provided. These indicators cover a comparative analysis of the planting sites and their surrounding environment, presence of other land cover within the reforestation site, and geometric characteristics of site boundaries, including nesting, intersection and administrative boundary alignments, and atypical geometric shapes. This approach will enable a comprehensive evaluation of spatial data quality that complements the ecological and socioeconomic success metrics currently used in restoration monitoring. Thus, the study's contributions are fourfold:

1. A consolidated dataset of global reforestation efforts at the planting site-level.
2. A location data integrity assessment of the georeferenced information based on a variety of indicators and data sources.

Indicator Name	Database Field Name	Definition	Data Completeness (%)
road_presence	total_road_length_km	Area within the site covered by roads	82
built_area_presence	built_area_2018	Area within the site covered by buildings or constructions	83
forest_at_planting_glad	treecover_atplanting	Area within the site covered by trees at planting time	99
other_landcover_score	other_land_cover_area_2020	Area within the site covered by other landcover e.g shrubs,grassland	98
nesting_polygon	nested_in	Checking if a site is contained inside another site fully	100
intersecting_Polygon	intersecting_with	checking if part of the site is overllaping with another site	100
exact_admin_area	exact_admin_area	Checking if the site is an exact representation of an administration area as defined by the GADM data	100
perfect_circle_indicator	polygon_circle_oval	Checking if the site is almost a perfect circle	100
geometry_validity	project_geometries_invalid	Checking for provided geometries if they represent a closed boundaries geometry or point geometry	100
stable_cropland_score	stable_cropland_cover_area_2020	Area covered by stable cropland from 2000 to 2020 within the site	87

Table 2. Location Data Integrity Score (LDIS) indicators. Data completeness indicates the percentage of non-missing entries for each indicator. All indicators use a binary (0/1) value range and equal weight (1).

3. Sentinel-2 satellite imagery time series for most planting sites in a convenient format to facilitate further analysis.
4. LLM-based text analysis of project description documents on contextual factors related to reforestation that are not well captured via remote sensing.

Besides descriptive analytics, this dataset could also be used as a weakly-labelled training dataset for a diverse set of image recognition tasks, e.g.to monitor longevity of reforestation efforts. Weak labels in this context are image attributes that are usually derived from secondary data sources and therefore might be subject to substantial levels of noise. Despite this lack of precision, weak labels can still serve as valuable signals for model training.

Methods

As illustrated in Fig. 1, we compile publicly available information on reforestation efforts from different websites, harmonize the available data and augment them with planting site-level information on the presence of roads and human settlements, on tree and land cover change and on weather characteristics, among other things. In addition, we utilize freely available Sentinel-2 imagery (specifically Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-1C) at 10 m resolution available from June 2015 onward to obtain relevant vegetation indices.

In the following sections, we describe the dataset generation in more detail.

Site Data Acquisition. To construct a global dataset of reforestation projects, we implement a systematic data collection process consisting of two main steps:

- **Identification of reforestation efforts:** In a first step, we conduct extensive open Google searches to identify potential websites of interest using the following keywords: "reforestation", "carbon registry", "tree planting", "climate mitigation tree" and "carbon market". Table 7 in the Appendix provides the list of websites selected to be of potential interest for our study.
- **Data retrieval:** The majority of websites do not provide project information in an easily retrievable format. While some websites such as Tree Nation (https://tree-nation.com) maintain well-documented Application Programming Interfaces (APIs), many websites display information via dashboards that do not allow for direct downloads and thus require scraping the relevant information. Table 7 provides details on the data collection mode and the reason why data collection was successful or not. It should be noted that a large share of websites of interest do not provide geographic information that allow pinpointing sites to a specific location. In most of these cases, the name of a larger administrative region is stated (e.g. "Yucatan, Mexico"), which, however, is insufficient to track reforestation progress via remote sensing. The retrievable location data are usually nested: A website hosts multiple reforestation projects which in turn are made up of multiple planting sites. We define a planting site as a closed, contiguous area with a non-zero area size. Consequently, if data are provided in complex polygonal structures, we break these up into individual polygons that adhere to our definition of a planting site. The data is stored at the site-level, where possible. During data retrieval, we capture metadata on project- and site-level, and on website-level as illustrated in Fig 2.

Site Preprocessing. As the data originates from different websites and is retrieved in different ways, data harmonization is needed, i.e. nomenclature and units of measurement need to be aligned. Furthermore, we

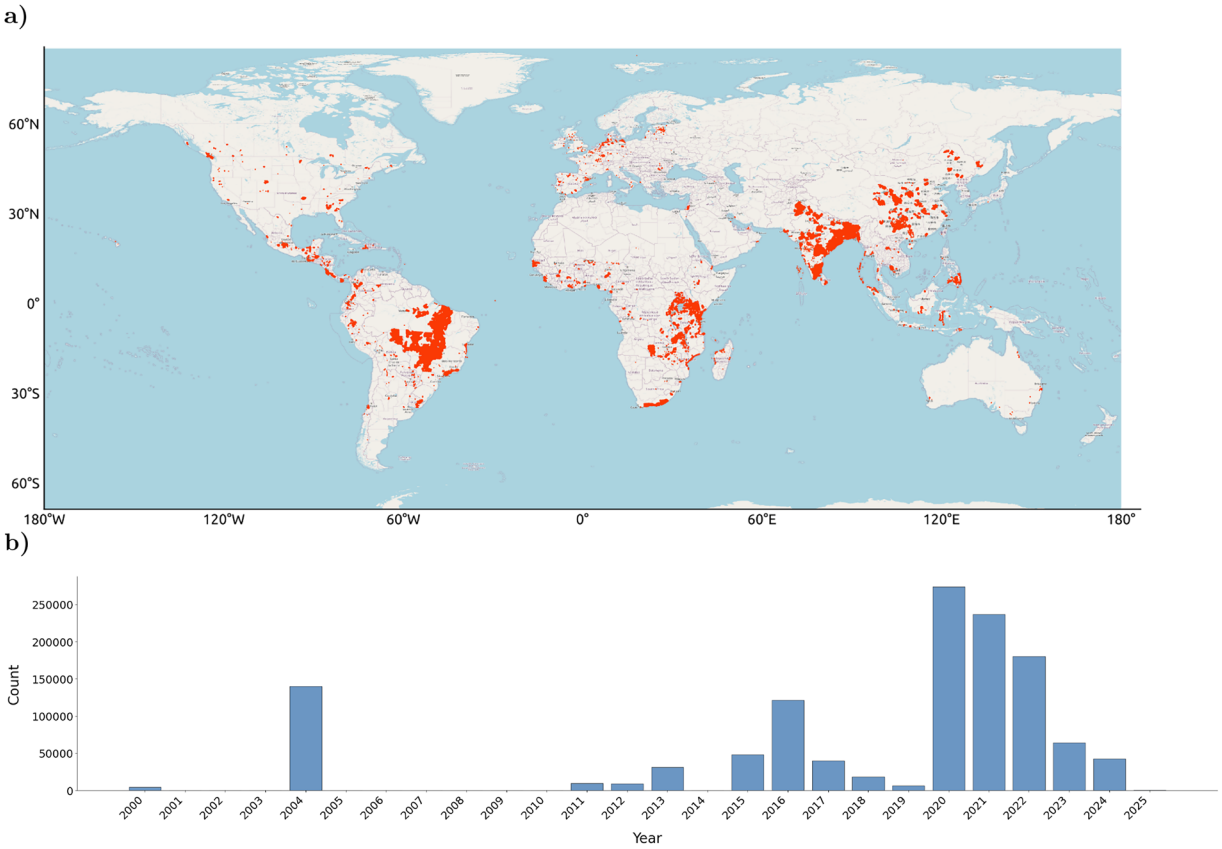

Fig. 2 Overview of reforestation "sites" in our dataset, by geographical location and planting date. (**a**) Reforestation site locations. Each red marker indicates the polygon boundary of a reforestation site (defined in our dataset as either a contiguous planting area or registered planted zone). (**b**) Count of reforestation sites by planting year. Each bar represents the number of distinct sites planted in a given year.

assume a considerable amount of imperfect duplicates in the data. There are multiple reasons for that: First, since the VCM lacks vertical integration, multiple actors may advertise the same project at various layers of the market. For example, the project developer wants to sell carbon credits, and therefore showcases their work to inform about their offers. At the same time, the certification body is interested in positioning itself as a trustworthy major player in the certification market and, therefore, wants to display all projects it certified. Simultaneously, companies that buy carbon credits from that project to offset their emissions are inclined to use it for communications purposes, such as corporate social responsibility campaigns. Second, the data are partly crowdsourced with little quality checks. Third, nesting might occur where individual sites are combined and displayed as a larger site by another website. Fourth, not all projects provide actual site locations in the form of areas covered. In many cases, either names of larger administrative regions are stated or point locations are provided. Both make it difficult to track reforestation efforts as the areas of interest are not clearly delineated. Specifically, the following preprocessing steps have been pursued to address some of these issues:

- **Aligning nomenclature:** Aligning column names and units of measurement (e.g. hectares vs km^2) and other naming conventions within columns is a prerequisite for cross-source comparisons. Indicators with information as reported on the website are denoted with the suffix *_reported*. Indicators estimated within this study are denoted with the suffix *_derived*. It should be noted that one key information, namely the planting date, is subject to some uncertainty as organizations report dates that may differ in their definition: some provide the project registration date, some the intervention year, some the crediting starting period and some the actual planting dates. We summarize the different definitions in *planting_date_reported* and give information about the definition type in the column *planting_date_type*.
- **Filtering for afforestation/reforestation projects:** Although our initial site-acquisition screening targets reforestation initiatives, further filtering is necessary because some of our data sources also include unrelated activities. To keep the scope strictly to afforestation and reforestation, we exclude projects outside these categories. We implement this by using existing project classifications where available, and otherwise by applying keyword matches to project names and descriptions.
- **Identifying special geometries:** Not all projects have polygonal site locations. However, clearly delineated areas are important for assessing and augmenting these sites. The same holds for site locations that are unrealistically large, e.g. spanning whole continents. Thus, we identify projects which closely resemble administrative areas (defined as overlapping each other with more than 98% of their respective area) according to the GADM

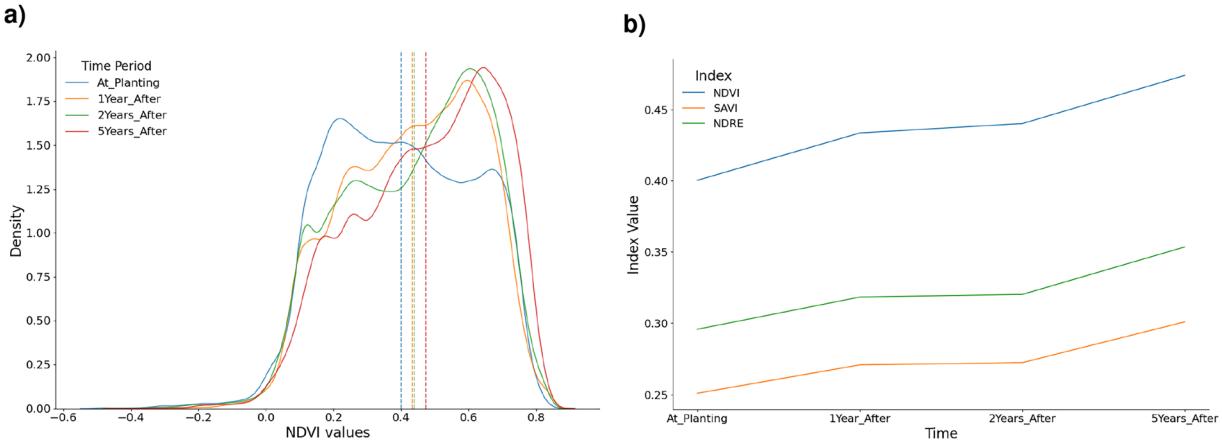

Fig. 3 Distribution of vegetation indices. (**a**) NDVI distribution. The normalized difference vegetation index (NDVI) values across reforestation sites, showing the variation in vegetation density for periods at planting, 1 year after, 2 years after and 5 years after planting. (**b**) Reforestation change using indices. Visualization of reforestation progress derived from vegetation index changes : Soil Adjusted Vegetation Index(SAVI), NDVI and Normalized Difference Red Edge index (NDRE) over time.

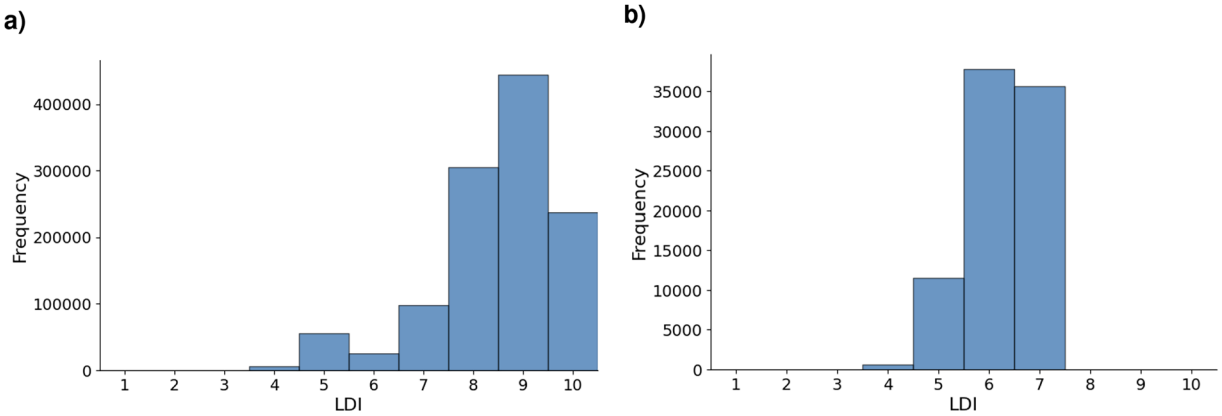

Fig. 4 Location data integrity score (LDIS) distribution, by geometry type. (**a**) Sites with provided area geometries. Shows the LDIS distribution for reforestation sites where detailed polygon geometries were available. (**b**) Sites with buffered point geometries. Displays the LDIS distribution for sites where only point locations were provided and area was inferred via buffering.

dataset[32] and store this information as the binary indicator *is_exact_administrative_area* in our dataset. We also calculate the circularity of each site stored as *circularity*. Geometries that closely resemble perfect circles (defined as a site being at least 98% circular) as near-perfect circularity hints at buffered (point) locations that most likely do not reflect the actual extent of a planting site. This information is stored again as a binary indicator called *is_perfectly_circular*. For sites described by a single point geometry we define a 100 meters buffer around each point and store this new geometry in the column *geometry_derived*. For polygon-based sites, we create a 500 m outer buffer, where the 'buffer' specifically refers to the annular region between the original polygon boundary and its 500 m extension; this allows us to compare within-polygon versus outside-polygon signals for boundary accuracy and to identify any spatial leakage of greenness measurements.

- **Identifying nested structures:** We note three different types of relationships between sites: subsets, duplicates, and supersets. If the area of the intersection of two sites constitutes more than 95% of the two individual areas, respectively, we assume that they are duplicates of each other. If the intersection constitutes more than 95% for one site, but less so for the other, we believe the former to be a subset of the latter and thus the latter to be a superset of the former. We, therefore, add the column *intersecting_with* to highlight the sites that intersect, and *nested_in* and *contains_small_polygon* to highlight subsets and supersets. This approach is crucial for avoiding double-counting of restoration areas and understanding the hierarchical organization of efforts. As shown in Table 4, nested structures are most prevalent among smaller sites (e.g., the substantial proportion of nested areas in sites under 500 km^2). In contrast, larger sites (above 1,000 km^2) are rarely nested in other polygons, suggesting most large-scale projects are independent and not embedded within even larger areas. Recognizing these nested relationships is essential for accurate spatial analysis and for assessing the true extent and distribution of reforestation activities.

Project Host	Projects	Sites	Georeferenced	Area (km^2)	Avg. Area per Site (km^2)
Verra	636	1,225,618	1,225,618	3.52×10^7	28.71
Gold Standard	71	—	—	—	—
Climate Action Reserve (CAR)	513	513	513	1,683	3.28
American Carbon Registry (ACR)	283	—	—	—	—
Other Hosts	44,316	69,648	63,668	6,579,569	103.34

Table 3. Project distribution across hosts. Number of reforestation projects and planting sites under major carbon-credit programmes. *Column definitions:* **Projects** = total number of reforestation projects; **Sites** = total number of discrete planting locations (- indicates data not available); **Georeferenced** = number of sites with valid spatial coordinates; **Area** (**km^2**) = cumulative georeferenced planting area in square kilometers; **Avg. Area per Site** (**km^2**) = mean area per georeferenced site.

Size Bin (km^2)	Count	Count (%)	Total Area (km^2)	Total (%)	Nested Area (km^2)	Nested (%)
< 10	1,225,267	99.68	53,188	1.17	6,858	12.89
10–50	1,423	0.12	28,885	0.63	4,302	14.89
50–100	232	0.02	16,500	0.36	2,799	16.96
100–500	636	0.05	156,765	3.44	24,145	15.40
500–1000	254	0.02	183,050	4.01	1,908	1.04
1000–2000	637	0.05	851,289	18.67	0	0.00
2000–5000	488	0.04	1,622,356	35.58	0	0.00
> 5000	237	0.02	1,648,053	36.14	0	0.00

Table 4. Distribution of reforestation site areas. Summary of site-area bins and their contributions. *Column definitions:* **Size Bin** (**km^2**) = range of site area; **Count** = number of sites within the bin; **Count** (%) = percentage of total sites; **Total Area** (**km^2**) = sum of areas of all sites in the bin; **Total** (%) = percentage of total area; **Nested Area** (**km^2**) = area of of sites fully inside other sites within the bin; **Nested** (%) = percentage of nested area relative to bin's total area.

Data Augmentation. To enhance the dataset's analytical depth, we integrate remote sensing data and secondary environmental indicators to evaluate the quality and impact of reforestation efforts. These augmentation steps address various aspects such as tree cover change, infrastructure presence, climate conditions, land use transitions, and topographic features. By incorporating these additional layers of secondary data, we aim to facilitate the cross-validation of self-reported project information.

Project Descriptions. Most sites or related projects provide a project description documenting project details. To extract species and planting-date information from these descriptions, we employ two transformer-based large language models:

- **BERT-Q&A:** We use the `bert-large-uncased-whole-word-masking-finetuned-squad` model[33], a large BERT model pre-trained with whole-word masking and fine-tuned on the squad benchmark, to extract the species planted.
- **DistilBERT-Q&A:** We use the `distilbert-base-cased-distilled-squad` model[34], a lighter, faster distilled version of BERT fine-tuned on squad, to extract planting dates.

These choices are motivated by the ability of these models to derive responses strictly from the provided context, thus avoiding hallucinations (i.e., cases where the model generates information not supported by the input) and improving reproducibility. Given the projects' respective descriptions, we ask them the following questions:

- What species were planted? Name each species mentioned.
- What is the planting date?
- What is the project start date?
- How many trees were planted?

Responses to each question are stored as separate indicators in the dataset. Some websites directly report the species planted, hence it is stored as the indicator *species_planted_reported*. In addition, we also extract information about the species planted from the project descriptions. We store this information as *species_planted_derived*. To ascertain the models' results, we manually checked 305 PDFs and compared them to the results provided by the models. The models correctly identified the species mentioned in 264 of the 305 PDFs and extracted accurate planting dates in 278 cases. When species were misidentified, the model typically returned general terms such as "native species," "indigenous tree species," "native vegetation types," or "locally adapted

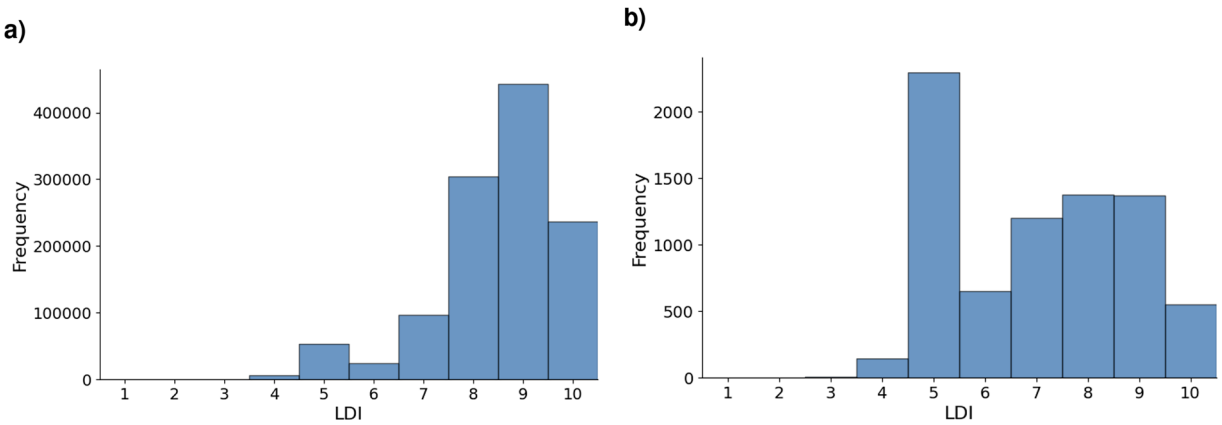

Fig. 5 Location data integrity score (LDIS) distribution, by area size. (**a**) Sites smaller than 5 km^2. Shows the LDIS distribution for reforestation sites with a reported or estimated area under 5 km^2. (**b**) Sites equal to or larger than 5 km^2. Displays the LDIS distribution for larger reforestation sites.

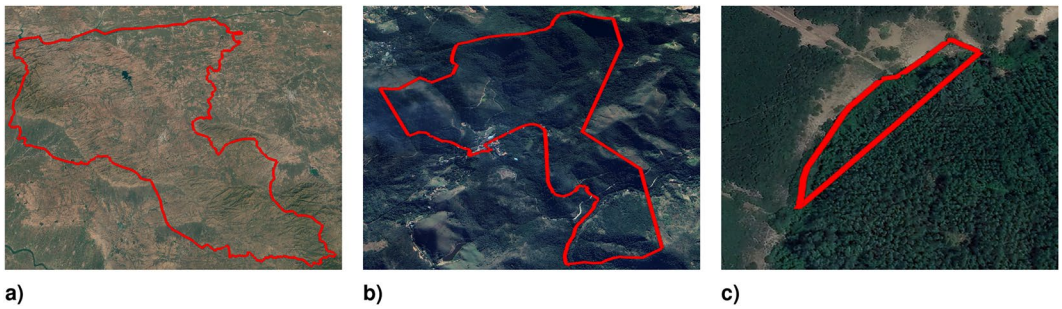

Fig. 6 Exemplary sites, by location data integrity score (LDIS). (**a**) Site with LDIS 2, area 5,438.7 km^2. (**b**) Site with LDIS 6, area 2.62 km^2. (**c**) Site with LDIS 9, area 0.00311 km^2.

species." Likewise, when planting dates were incorrect, the model often produced broad timeframes (e.g., "minimum 40 years," "4-6 years") or general targets (e.g., "by 2030").

Tree Cover Loss. We also consider tree cover loss to be an important indicator for monitoring reforestation efforts, particularly with respect to their additionality and permanence. Have the sites experienced measurable tree cover loss prior to planting? Have the sites experienced tree cover loss in the years following planting?

In our study, we assess tree cover loss patterns using the Global Forest Watch 2023 v1 dataset[21]. This dataset detects forest loss events at a minimum mapping unit of 30 m resolution per pixel. Tree cover loss is defined as a stand-replacement disturbance or complete removal of the tree canopy. Also it does not distinguish between permanent deforestation and temporary losses (e.g., due to fire or logging followed by re-growth). We use this dataset to estimate the proportion of tree cover loss within each site for the five years before planting, the year immediately preceding planting, and the five years after planting, where applicable, for the period 2000 to 2023.

Infrastructure Presence. The extent of human presence within reforestation sites may provide information on the accuracy of the sites' reported geographical boundaries: if the supposed planting site overlaps with a human settlement, then the site's boundaries are unlikely to be accurate. Consequently, we report the proportion of area under human settlements within each site using the GHS-BUILT-S dataset from the P2023 release of the Copernicus' Global Human Settlement Layer (GHSL)[35]. Furthermore, we use NASA's gROADSv1[36] dataset to estimate the length of roads (in km) per square kilometer within a given site.

Climate Features. Climatic conditions such as precipitation and temperature have a strong influence on forest growth and reforestation success[37,38]. Using the historic WorldClim API[39], we extract the climate features, i.e. average monthly rainfall and minimum and maximum temperature experienced within each reforestation site at different points in time: at planting date and one year, two years and five years after planting using the site midpoint geometries.

Site Terrain. Furthermore, we include the elevation and slope variables derived from the Shuttle Radar Topography Mission (SRTM)[40–42] For each site, we extract all intersecting $90m \times 90m$ raster cells to compute mean elevation and mean slope indicators. However, we note that SRTM's 90 m resolution may not capture very fine-scale topographic conditions.

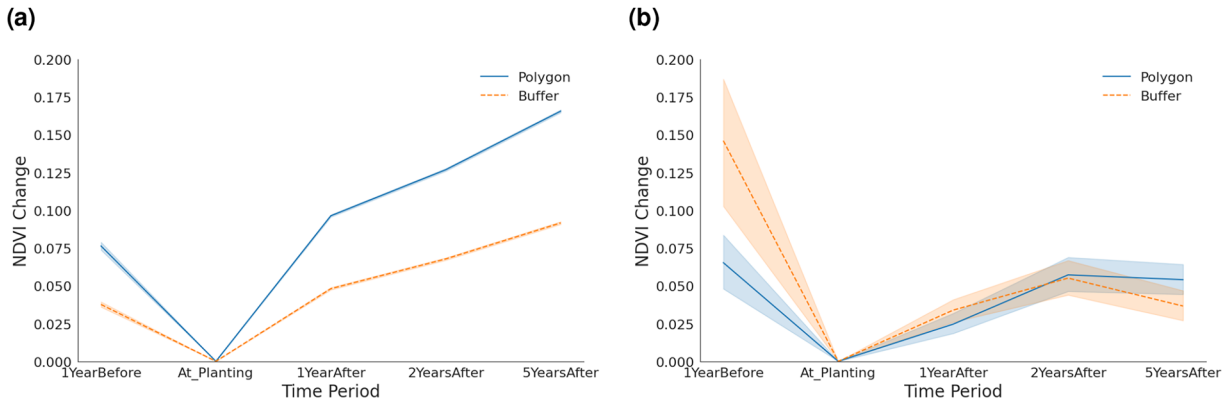

Fig. 7 NDVI change over time. Comparing planting sites and their buffer areas, by area size. Uncertainty is represented by bootstrapped 95% confidence intervals. (**a**) Sites smaller than 5 km^2. (**b**) Sites equal to or larger than 5 km^2.

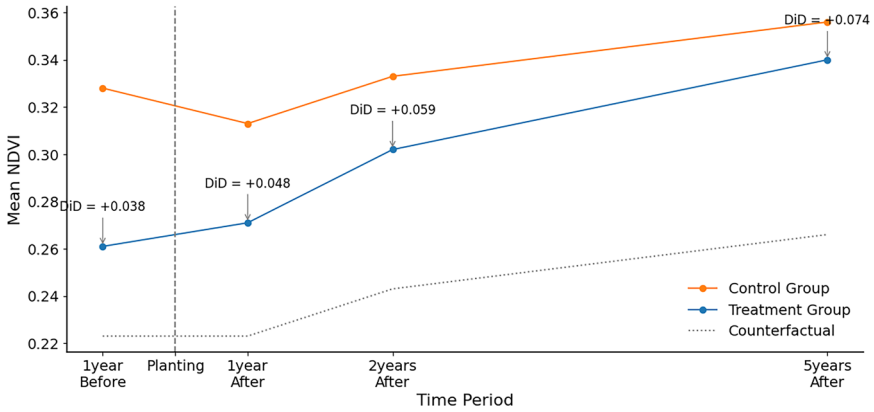

Fig. 8 Evolution of the NDVI over time. Subset of sites with NDVI data for all relevant dates around planting. NDVI averaged across sites. The control group is composed of buffer areas (defined by a 500 m buffer around the planting site) around respective planting sites. DiD values are estimated against planting date NDVI values and correspond to the difference in NDVI between the Treatment group and the Counterfactual. The counterfactual are model-based NDVI predictions excluding the interaction (*DiD*) component.

Land Cover and Land Use. Land use conflicts are prevalent and a major driver of deforestation[43]. Identifying areas where reforestation would directly compete with existing livelihoods can help prevent potential conflicts with local communities that depend on these resources, which could lead to the failure of these reforestation projects. Therefore, understanding the prevalence of different land cover types at project sites is a potentially relevant indicator for reforestation success. To unveil land cover change patterns within the project area at various points in time, we use the Global 2000-2020 land cover and land use change (GLAD) dataset[23] to extract proportions of reforestation site area with following land cover transitions: cropland from tree, cropland to tree, short vegetation after tree loss, stable cropland 2000-2020 and permanent water. While the GLAD-derived Global Forest Watch layer provides annual, wall-to-wall coverage, its classification accuracy varies by biome and transition type. For instance, the tropical-forest gain class exhibits a false-negative rate of approximately 52%, meaning over half of real gain events in tropical regions may go undetected. Overall, global false-positive rates for forest loss and gain are roughly 13% and 23.6%, with corresponding false-negative rates of 12.2 % and 26.1 %[21]. These limitations imply that our estimates of both loss and gain could under- or over-represent true changes, and thus should be interpreted with appropriate caution.

Annual Tree Cover Change Indicators. Existing datasets such as[21] and[23] do not provide tree cover extent and gain estimates on an annual basis. The Normalized Difference Vegetation Index (NDVI)[44] is widely used to monitor reforestation and vegetation changes.[45,46] have shown its effectiveness in assessing reforestation success and tracking changes in vegetation cover over time. According to[47], the NDVI is positively correlated with standing stock, particularly on larger spatial scales. However,[45] indicates that NDVI alone may not always capture substantial changes in forest cover, suggesting the need for integrated monitoring approaches. According to[46], combining NDVI with other indices, such as the Normalized Difference Red Edge (NDRE) index, can provide a more comprehensive assessment of afforestation and reforestation dynamics. Thus, as vegetation indices of

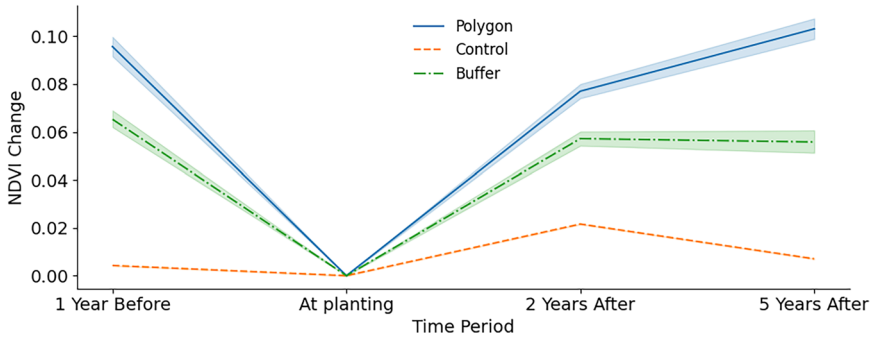

Fig. 9 NDVI change over time in Kenya. Comparison of planting sites, their buffer areas, and corresponding synthetic control groups. All sites are smaller than 5 km^2.

choice, we select the NDVI as a measure for presence or absence of tree cover, the NDRE for measuring a vegetation's health and the Soil-Adjusted Vegetation Index (SAVI) to correct the NDVI from any soil reflectance when the vegetation cover is sparse (e.g. when trees are still young and canopy cover sparse). All these indices are calculated from Sentinel-2 imagery COPERNICUS/S2_HARMONIZED" as we aim to cover most of our reforestation sites from their earliest dates. Table 1 gives a summary of the definitions of the three indices used.

To limit the impact of seasonal fluctuations on our vegetation indices, we calculate annual vegetation indices based on the top three greenest months as measured by monthly NDVI averages in 2023 (in Africa: Jan, Apr, May, in Asia: Jun, Jul, Aug, in Europe: Apr, May, Jun, in South America: Feb, Mar, Oct). As illustrated in Fig. 3, the NDVI changes by years around planting, but behaves similarly to the NDRE index and the SAVI.

The NDVI distribution shifts progressively to the right—from a mean of 0.39 at the (likely) planting date to 0.47 five years later, indicating a clear increase in vegetation greenness. The narrowing of the distribution and the higher peak density at year 5 suggest not only recovery but also greater uniformity in canopy cover. Interestingly, a subset of sites already exhibited NDVI values above 0.60 at planting, highlighting that some "reforestation" areas in our database were dense forest from the outset. This overall trend confirms NDVI as a plausible and sensitive proxy for monitoring post-planting. However, we also consider the NDVI, NDRE and SAVI as our main indicators of reforestation success over time as local variations between the indicators may apply.

Location Data Integrity Assessment. The accuracy of the provided delineations of the planting sites is a crucial element for correctly assessing and quantifying reforestation efforts from space. However, even planting sites with perfectly delineated polygons do not guarantee successful revegetation without proper planting protocols, periodic monitoring, and attention to surrounding land-use activities. We try to understand a site's geographic integrity by triangulating different perspectives:

1. We check for other human infrastructure or water bodies within the planting area. If built-up areas, roads, or water bodies cover more than 10% of the planting site, we assume that the delineation of the actual planting area lacks accuracy.
2. We consider land cover classifications and land cover conversions after planting. If other land cover classes are present and cover 20% or more of the planting area, we take this as an indication that the planting is not accurately delineated.
3. We account for possible double counting in projects by checking the intersecting (*intersecting_with*) and nested polygons (*nested_in*). If the planting areas are nested within or intersecting with each other, we take this as an indication of inaccurate delineation and/or potential double counting.
4. We check whether the site boundaries resemble (subnational) administrative areas available through GADM[32]. Named *exact_admin_area*, this binary indicator takes a value of 1 if a site shares more than 98% of its area with an administrative area and vice versa, 0 otherwise.
5. We assess if the provided site boundaries are perfectly circular, hinting at inaccurate actual planting site boundaries, e.g. by approximating them via buffered point locations. Again stored as a binary indicator, *polygon_circle_oval* takes the value of 1 if the site boundaries are 95% circular, 0 otherwise.
6. The invalid geometries indicator *project_geometries_invalid* is used to identify whether the reported geometries are valid. Specifically, this check ensures that the provided geometry is a properly closed polygon, which is necessary for consistent spatial analysis and monitoring.
7. We check for the presence of forest on the planting date. If the forest at the planting covers 20% of the planting site area or more, we take this as an indicator of the planting site not accurately delineated.
8. We check for the presence of stable cropland area within the planting site between the years 2000 and 2020. If stable cropland covers 20% or more of the planting area, we take this as an indication that the planting is not accurately delineated

Most of the indicators require an area to be calculated. Since some planting sites are originally just provided as point locations, we use their derived geometry (i.e. buffering a site's point location with a 100m buffer) for further analysis. However, this naive approach most likely does not capture the true extent of a planting site. Thus, we expect the LDI scores for this group of sites to be lower on average than for planting sites that come

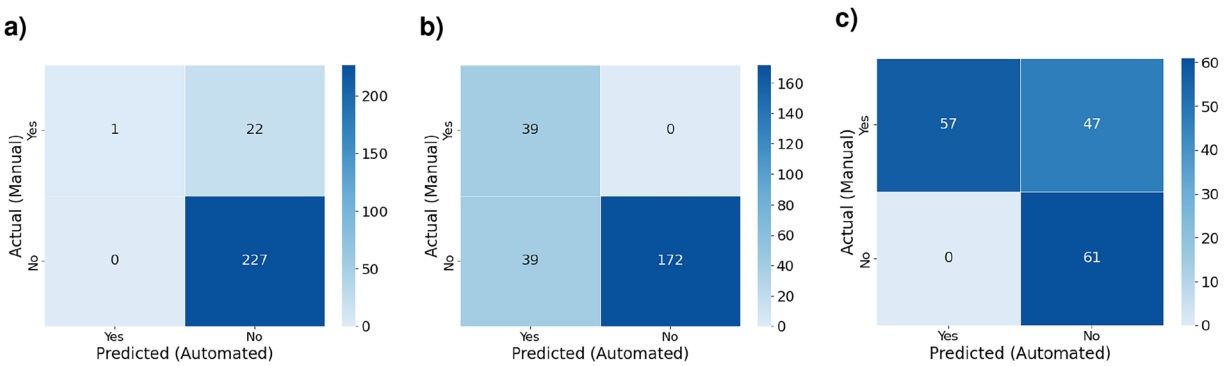

Fig. 10 Comparing derived indicators with manually labelled annotations. (**a**) Presence of roads. (**b**) Presence of built-up areas. (**c**) Forest at planting.

Geometry Capturing Planting Area			
LDI Score	**No**	**Yes**	**Total**
6 and below	87.5%	12.5%	8
7–8	10.2%	89.8%	128
9	4.4%	95.6%	114
Total	25	225	250

Table 5. Comparing location data integrity score (LDIS) against manually annnotated images.

with a delineated area. Figure 4 shows the LDIS distribution for planting sites in the dataset for those two groups separately.

For the group of reforestation sites with area geometries (Fig. 4a), 79% of these sites have an LDIS of 9 and less, implying that only 21% sites meet all our quality indicators. For the group of sites with buffered point geometries (Fig. 4b), the number of projects that score perfectly on the LDI assessment reduces to 0%.

Data Records

The dataset associated with our study can be found in a public data repository[48]: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZJODGO. It contains georeferenced information on reforestation efforts around the globe, including hierarchically-organized metadata. Information are stored on the level of individual reforestation sites, where possible. If site-specific data is not available – for example, when geographic information are not provided as actual planting sites, but via higher-level location names – information are stored at the project- or host-level. Site data is further augmented with indicators on tree cover change, forest loss, land cover change, and the presence of human settlements and infrastructure within these reforestation sites, among others. A detailed description of each indicator in our dataset can be found in Tables 8, 9 and LDI indicators extracted from the variables in Table 2. The dataset spans approximately 33 years and is based on the analysis of nearly 50 organizations working in the reforestation space. Links to the sources are provided both in Table 7 as well as in the column *url* of the dataset. Primary datasets from these websites are available upon request or retrieval can be repeated using the code published alongside the dataset. All the data records are consolidated into a single geographic Parquet file (`.parquet` format) and are also provided in CSV format (`.csv`) for broader accessibility. Due to repository upload requirements, the CSV files are split into four parts corresponding to the categories *Old World Part 1*, *Old World Part 2*, *New World*, and *Antarctica*. Alongside the georeferenced dataset, we download Sentinel-2 median composite scenes for each planting site, even those lacking pre-computed cloud flags, and for each scene, compute the within polygon cloud fraction using the QA60 band. We then retain only those composites where cloud cover over the site footprint is less than 20 %. These images correspond to four distinct time points: the planting year, one year before planting and one, two, and five years after planting. For each of these periods, we selected the composite image from the month with the highest Normalized Difference Vegetation Index (NDVI). We used these composite images to calculate NDVI, SAVI, and NDRE, as analyzed and discussed in section "Annual Tree Cover Change Indicators" above, by focusing on those months with top NDVI values. The images are provided as TIFF files containing georeferenced remote sensing imagery for individual reforestation sites. All image files follow a consistent naming convention to ensure clarity and reproducibility. Each file name is structured as:

```
[site_id_created]_[time_period]_[year]_[bands].tif
```

Here, **site_id_created** is the unique numeric identifier for each site (e.g., `1000`); **time_period** indicates the monitoring stage, recorded as `atplanting`, `1yr_after`, `2yr_after`, or `5yr_after`; **year** is the calendar year in which the image was collected (e.g., `2019`); and **bands** refers to the image type or spectral bands (e.g., `RGB`).

	Dependent variable: NDVI			
	One Year Before	**One Year After**	**Two Years After**	**Five Years After**
	(1)	**(2)**	**(3)**	**(4)**
Intercept	0.290^{***}	0.265^{***}	0.265^{***}	0.265^{***}
	(0.001)	(0.000)	(0.001)	(0.001)
treat (g)	-0.105^{***}	-0.090^{***}	-0.090^{***}	-0.090^{***}
	(0.002)	(0.001)	(0.001)	(0.001)
time (t)	0.038^{***}	0.048^{***}	0.068^{***}	0.091^{***}
	(0.002)	(0.001)	(0.001)	(0.001)
treat*time (gt)	0.038^{***}	0.048^{***}	0.059^{***}	0.074^{***}
	(0.002)	(0.001)	(0.001)	(0.001)
Observations	139250	469500	469500	469500
R^2	0.061	0.081	0.097	0.132
Adjusted R^2	0.061	0.081	0.097	0.132
Residual Std. Error	0.208 (df=139246)	0.170 (df=469496)	0.181 (df=469496)	0.184 (df=469496)
F Statistic	2995.626^{***} (df=3; 139246)	13736.170^{***} (df=3; 469496)	16738.387^{***} (df=3; 469496)	23880.764^{***} (df=3; 469496)

Table 6. Difference-in-Differences regression results for NDVI. *Note:* $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

For example, the file `1000_1yr_after_2019_RGB.tif` corresponds to site 1000, imaged One year after planting, data were collected in 2019 using RGB bands.

Technical Validation

The purpose of this dataset is to consolidate publicly accessible information on reforestation efforts, augment these with indicators derived from metadata and secondary data and assess the integrity of the location data provided for these efforts on the level of individual planting sites. The dataset sets out to be comprehensive in scope and instead of excluding sites from the beginning, we aim to flag data quality constraints through a well-defined selection of indicators. To validate the scope of projects included in the dataset, we compare it with the Voluntary Registry Offsets (VRO) Database of the Berkeley Carbon Trading Project (https://gspp.berkeley.edu/research-and-impact/centers/cepp/projects/berkeley-carbon-trading-project/offsets-database), which contains an updated list all carbon offset projects from the registries of the four major independent carbon credit programmes (namely Verra, Gold Standard, Climate Action Reserve and ACR). These four programmes account for the vast majority of the global VCM[49]. We find that we capture all projects that are also in the VRO database, which make up 85% of all the area covered in our dataset (see Table 3). Table 3 shows how carbon-credit projects are distributed across the world's leading registries. "Projects" is the number of distinct carbon-credit projects registered under each host. "Sites" is the total number of individual planting polygons reported by those projects (if a host does not publish polygon data, its entry is "-"). "Georeferenced" counts only the sites with valid latitude and longitude coordinates. "Area" is the sum of all georeferenced site polygons expressed in square kilometres, and "Avg. Area per Site" is the mean polygon size for those georeferenced sites. Together, these metrics reveal both the scale (total area) and granularity (number of sites) of each registry's portfolio. As both our database as well as the VRO database store the original unique project identifiers of the respective carbon credit programmes (in our case *project_id_reported*), the databases can easily be matched for further analysis.

Validating how many of all reforestation projects in the world we do not capture, proves more challenging. During data collection, we observed that publicly accessible information on government-led projects is especially sparse. As around 70% of global forestland is under "legal and administrative authority" of governments[50], we assume that government-led reforestation efforts are under-represented in the current dataset.

On the level of planting sites, we observe that the 1% largest sites make up about 90% of all the total reforested area in our database (see Table 4). Table 4 summarizes the distribution of reforestation site areas. "Size Bin" refers to the range of individual site areas (in km^2). "Count" is the number of sites that fall into each size bin, and "Count (%)" shows what percentage of all sites that represents. "Total Area" is the cumulative area of all sites in each bin, and "Total (%)" reflects the share of the total reforested area. "Nested Area" refers to overlapping areas within larger planting polygons, and "Nested (%)" is the proportion of bin area that is nested.

By disaggregating the LDI score by area size, we observe that larger sites score considerably lower than smaller sites (see Fig. 5).

Looking at the underlying LDIS indicators for the largest sites, we further observe that most of these sites closely resemble an administrative area or are almost perfectly circular. In those cases, the likelihood that infrastructure or other land cover areas are also covered by a given site is quite high, thus explaining the low LDI scores. Figure 6 provides examples of what the sites for a given LDIS look like.

For a subset of 250 manually annotated satellite images from 250 randomly selected reforestation sites, we capture the binary variable *Geometry_Capturing_Planting_Area*, which is *yes*, if the reforestation area is accurately and comprehensively captured without other activities visible on-site (see Table 5).

At LDIS of 9, for 95.6% of the sites there is agreement with the manually annotated cases which is an increase from the 12.5% and 89.8% for an LDIS of 6 and below and between 7 and 8, respectively. To further understand the quality of the planting site boundaries exactness and any leakages, we set a 500 m buffer around each

planting site polygon provided. Figure 7 shows the mean NDVI change per period across planting sites and their respective buffer areas, with a 95% bootstrap confidence interval around it.

Although the NDVI changes of the buffer and planting sites in our data set are notably different, both show a similar increase over time. Potential reasons for that are spillover and edge effects and growth of shrubs and possible presence of mature trees around the site the buffer. Similar cases have been witnessed in other studies such as[51] and[52] monitoring the buffer and planting using vegetation indices.

Interestingly, we also observe considerably higher NDVI values in the year pre-planting compared to the planting year. Potential reasons for this are multiple such as clearing sites from shrubs, grass and weeds before planting or reforestation efforts being a consequence of previous wildfires or other adverse events. Manually inspecting satellite imagery for a random subset of planting sites that show significant NDVI drop from the pre-planting to the planting year does not reveal obvious common causes for this phenomenon as it ranges from shrubs and grass land before planting to other cases having forest before planting. However, we consider this phenomenon (including land management activities like tillage and land preparations before planting[53,54]) to be relevant especially in the context of additionality and permanence of reforestation efforts and therefore regard further investigations in that direction as an excellent use case of our database. Moreover, in Fig. 7a, the CI around the mean NDVI change line is very narrow, indicating more precise estimates at smaller sites. By contrast, Fig. 7b shows wider shaded bands for larger polygons, reflecting greater variability in mean NDVI change in larger polygons in our database.

To investigate the significance of the planting effect on the NDVI more rigorously, we apply a Difference-in-Difference (DiD) approach. In the DiD design, we consider planting as treatment with NDVI values before and after treatment within and outside the planting area (defined by a 500 m buffer around the planting site) as the control, respectively, as our outcome of interest. We subset our dataset to sites that have NDVI observations available for every pre-, at and post-planting period (excluding 5-years before planting due to data sparsity), thereby ensuring a balanced panel.While we recognize that land-cover histories may differ — with planting areas often being previously deforested or degraded, and buffers potentially retaining existing vegetation or representing recovering secondary forests — we assume the common trend assumption (i.e. vegetation growth within and around the planting area would be similar if trees had not been planted) to hold. This assumption is supported by the preceding analysis (see Fig. 7, which shows that NDVI change trends between planting areas and their buffers are closely aligned not only prior to planting but also in the years immediately following it.

The DiD approach is implemented as a linear regression with a time-treatment interaction effect (see Equation 1). The general form of the DiD model is:

$$\mathrm{Y}_{it} = \text{intercept} + \beta_1 \cdot \mathrm{g}_i + \beta_2 \cdot \mathrm{t}_j + \beta_3 \cdot (\mathrm{g}_i \times \mathrm{t}_j) + \epsilon_{it}, \tag{1}$$

where Y_{it} represents the outcome of interest, namely the average NDVI at the planting site i at time t. g_i is a binary variable that indicates whether it is a treatment (planting site) or a control (buffer), t_j indicates time as either before or after the planting date, and ϵ_{it} is the error term. The interaction term $\mathrm{g}_i \times \mathrm{t}_j$ describes the effect of interest, for which β_3 captures the effect of treatment (the DiD estimate). The results of the DiD approach are shown in Table 6.

The DiD regression results in Table 6 and Fig. 8, confirm a statistically significant positive impact of reforestation as measured by the NDVI one year before and one, two and five years after planting. Interestingly, the NDVI values of buffer areas are higher than in their respective planting sites with the difference growing around as presented by the negative coefficients of the treatment term g in regression (1) and (2), thus hinting at site clearing efforts at planting sites. This is supported by the positive time term t before planting. After planting, it takes more than two years of tree growth to compensate for the different offset of buffer and planting sites as shown by the difference between the treatment term g and the planting effect gt. The positive time term t after planting implies that no extensive logging or clearing has taken place since planting.

To validate that these temporal changes especially in the buffer areas are site-specific trends and not driven by higher-level dynamics such as climate change, we apply the synthetic control method to identify the "counterfactual" trend. Although we perform this validation step exemplary for one country, namely Kenya, we assume that the results are by-and-large representative for other regions in the world. To create a synthetic control group to our planting sites (in this context we consider "planting" as the treatment), we randomly select 10,000 points in Kenya, calculate their NDVI at planting and allocate each of the values to a bucket derived from NDVI distribution of the reforestation sites. We then use weights from these NDVI distributions to calculate the weighted mean across the buckets, resulting in NDVI averages for the synthetic control group. Figure 9 compares the NDVI changes over time for the planting sites, their respective buffer areas and their synthetic controls for sites located in Kenya.

Similar to the NDVI changes over time presented in Fig. 9 for all sites, the Kenyan sites show similar trends for the planting sites and their buffer areas. However, the synthetic control group remains largely flat as one would expect in the absence of common dynamics. This validates our initial assumptions that a) noticeable site-specific planting effects exist, and b) the sites are to a large extent not properly delineated or other mechanisms cause the buffer area to behave similar to the corresponding planting site.

In a final step, we validate our indicators derived from secondary data by the randomly sampled 250 sites with annotated historic high-resolution images of these sites accessed via Google Earth Pro and Global Forest Watch (GFW) Map before, at and after planting. We check for the presence of roads, buildings, and forests of each reforestation site in a binary manner. We compare our manual annotations with indicators using a confusion matrix and calculate accuracy and F1 score as illustrated in Fig. 10.

For the presence of roads within the sampled planting sites, we calculate an accuracy of 91% precision and an F1 score of 0.87, effectively identifying roads compared to manual annotations, with minor discrepancies in

recognizing smaller roads. For the presence of built areas, we calculate an accuracy of 84% and an F1 Score of 0.86, indicating good precision in mapping built areas. Differences arise due to varying definitions of built-up areas, as some open spaces and roads are classified as built areas in the automatic framework using the GHS BUILT settlement data. For the presence of forests at the time of planting, we compare our indicators to manual annotations using high-resolution images from Google Earth Pro and GFW, as well as lower-resolution Sentinel-2 and Planetscope images for cases where there are no timely high-resolution images. For detecting the presence of forests, we calculate an accuracy of 71.51% (F1 Score: 0.71), thus performing moderately well. A potential explanation with the comparatively low alignment of manual annotations and dataset indicators can be found in the varying definitions of what classifies as a forest.

Usage Notes

This dataset sets out to provide perspectives on reforestation success of reforestation efforts around the globe. Specifically, it provides an assessment of the integrity of the location data provided for each reforestation site – an essential pillar for any downstream remote monitoring effort. However, the impact of these reforestation efforts is multi-dimensional. The dataset barely considers ecological, legal, managerial and socio-economic perspectives and consequently cannot be used to assess these dimensions and/or the overall quality of a given reforestation project.

Although we made efforts for this dataset to be as comprehensive as possible, there are multiple reasons why some reforestation efforts are not included: First, reforestation efforts without any publicly available information on the internet are excluded. While this might hold true for only a tiny fraction of projects that aim to participate in the VCM, we got the impression that especially government-led planting initiatives are much less well documented publicly. Second, since the data collection is based on keyword-specific Google searches, we might miss out on some initiatives that do not align with the chosen keywords. While we are confident that this is a minor issue for projects that host information in English, we might miss out more substantial parts of the reforestation space that provides information only in languages other than English. However, since the VCM, especially the carbon credit certification part, is dominated by organizations registered in the US and Switzerland, we expect that most projects that plan to participate in the VCM will have information in English listed. Third, some websites mention active reforestation efforts, but provide too little detail or data in machine-readable format (e.g. providing maps of site locations without annotations in .pdf-format) to add it to the dataset. Since we expect this to apply to a non-negligible share of projects, we flag those websites in Table 7 accordingly. Further, in order to use the dataset effectively, consider the following:

- Software for analysis: The dataset is compatible with various geographic information system (GIS) software, including QGIS, ArcGIS, and Google Earth Engine. It can also be used with most geospatial processing libraries of popular programming languages such as Python and R.
- Metadata interpretation: Both geographic data and project descriptions in the dataset are taken *as is* from the respective websites. We do not correct any of these information, but use secondary data to assess their validity. Refer to Tables 2, 8 and 9 for a detailed description of each indicator and its source.
- Hierarchical data: The dataset provides information on the level of planting sites. Each planting site has its own identifier assigned (*site_id_reported*). Most sites are part of a reforestation project. A reforestation project usually consists of multiple planting sites and has its own identifier (*project_id_reported*). While some reforestation projects may stretch across multiple countries, they are usually confined within the national boundaries of one country, denoted by its ISO-3 country code (*iso3*). In addition, most websites host more than one reforestation project. Websites are listed with their *url* and the website's name (*host_name*). The hierarchical structure of the data and the cross-level relationships need to be taken into account when conducting any analysis on the dataset.
- Planting date: As already noted in Section, there is considerable uncertainty related to the indicator *planting_date_reported* as the underlying definitions planting date may vary across hosts. Please consider the indicator *planting_date_type* and potential definitions from the hosts' websites for more details.
- Temporal analysis: The dataset enables temporal analysis of reforestation sites. However, although we use the same data sources across years and pay attention to flag methodological changes that may limit comparability over time, undeclared methodological changes may still occur. Users can track changes over time by comparing tree cover and vegetation-related indices (e.g. NDVI) at, one year after, two years after and five years after the planting date, if applicable. Tree loss and land cover conversion indicators are also available for the years prior to planting.
- Integration with other datasets: Users may integrate this dataset with other environmental datasets for a more comprehensive analysis of reforestation's impact on biodiversity, land use, and climate change. This can be mainly done via the geographic location of the planting sites (*geometry_reported*). Alternatively, since we report the original identifiers both for the sites and the projects, if available, they, too, can be used to integrate site- and project-specific information.
- Caution with interpretation: While the dataset provides a standardized assessment of the location data integrity, users should exercise caution when drawing conclusions about the overall reforestation success as reforestation success can be influenced by various factors not captured by the dataset as mentioned before.
- Data retrieval and re-use: The dataset is intended for scientific use and for supporting monitoring and verification processes in the carbon markets. While some programmes such as the ESA Copernicus Sentinel 2 (https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_HARMONIZED) explicitly allow for reuse, the terms and conditions of use may vary across websites.

Uncertainty. The dataset[48],relies on reported data and a variety of highly processed secondary data sources that both come with their own intricate error frameworks. It is beyond the scope of the paper to quantify and integrate the various sources of uncertainty that are inherent to this dataset. Just to name a few: Geographic boundaries of planting sites are prone to data entry errors and likely depend on the accuracy of the GPS device. In NASA's gROADS dataset, the underlying road representation dates range from the 1980s to 2010 and originate from different national sources with varying data qualities. Global land cover datasets exhibit considerable uncertainties[55,56]. These limitations have to be especially considered for any future uses that rely on the correct approximation of the underlying error structure, e.g. for hypothesis testing.

Data availability

A Global Dataset of Location Data Integrity-Assessed Reforestation Efforts, the dataset associated with this study is publicly available at the Harvard Dataverse[48]: https://doi.org/10.7910/DVN/ZJODGO. The data are provided in Parquet, CSV, and TIFF file formats. Additional details on the dataset contents, versions, and variables are described in the Data Records section.

Code availability

The programs used to generate the dataset are Python 3.10 and Google Earth Engine (GEE). The code and instructions to reproduce the dataset are available on https://github.com/Societal-Computing/Forest_Monitoring.

Received: 19 May 2025; Accepted: 2 September 2025;
Published online: 29 October 2025

References

1. Food and Agriculture Organization of the United Nations (FAO). *Global Forest Resources Assessment 2020*. Food and Agriculture Organization of the United Nations, 2020. http://www.fao.org/forest-resources-assessment/en/.
2. Rytter, R.-M. & Rytter, L. Carbon sequestration at land use conversion–early changes in total carbon stocks for six tree species grown on former agricultural land. *Forest Ecology and Management* **466**, 118129 (2020).
3. Alados, C. Two dimensional searching paths exhibit fractal distribution that change with food availability (Normalized Difference Infrared Index, NDII). *Ecological Indicators* **139**, 108940 (2022).
4. Russo Lopes, G. & Bastos Lima, M. G. Understanding deforestation lock-in: Insights from Land Reform settlements in the Brazilian Amazon. *Frontiers in Forests and Global Change* **5**, 951290 (2022).
5. West, T. A. P., Börner, J., Sills, E. O. & Kontoleon, A. Overstated carbon emission reductions from voluntary REDD+ projects in the Brazilian Amazon. *Proceedings of the National Academy of Sciences* **117**, 24188–24194, https://doi.org/10.1073/pnas.2004334117 (2020).
6. West, T. A. P. *et al*. Action needed to make carbon offsets from forest conservation work for climate change mitigation. *Science* **381**, 873, https://doi.org/10.1126/science.ade3535 (2023).
7. Ecosystem Marketplace. Voluntary Carbon Markets Rocket in 2021, On Track to Break $1B for First Time. *Ecosystem Marketplace* https://www.ecosystemmarketplace.com/articles/press-release-voluntary-carbon-markets-rocket-in-2021-on-track-to-break-1b-for-first-time/ (2021).
8. cCarbon.info. Uncertainty in the VCM III: The Future of REDD+ in the VCM. *cCarbon* https://www.ccarbon.info/article/uncertainty-in-the-vcm-iii-the-future-of-redd-in-the-vcm/ (2023).
9. Carbon Pulse. REDD+ Market Slump and Recovery. *Carbon Pulse* https://carbon-pulse.com/191565/ (2023).
10. Probst, B. S., Toetzke, M., Kontoleon, A. & Díaz Anadón, L. Systematic assessment of the achieved emission reductions of carbon crediting projects. *Nature communications* **15**, 9562 (2024).
11. Food and Agriculture Organization of the United Nations. *FRA 2000: Global Forest Resources Assessments*. Food and Agriculture Organization of the United Nations [Internet], https://www.fao.org/3/y1997e/y1997e00.htm (2000).
12. Sexton, J. O. *et al*. Conservation policy and the measurement of forests. *Nature Climate Change* **6**(2), 192–196 (2016).
13. Stockholm Environment Institute and Greenhouse Gas Management Institute. Carbon Offset Guide http://www.offsetguide.org/wp-content/uploads/2020/03/Carbon-Offset-Guide_3122020.pdf (2020).
14. Prussi, M. *et al*. CORSIA: The first internationally adopted approach to calculate life-cycle GHG emissions for aviation fuels. *Renewable and Sustainable Energy Reviews* **150**, 111398 (2021).
15. Integrity Council for the Voluntary Carbon Market (ICVCM). Core Carbon Principles (2024).
16. Martin, M. P. *et al*. People plant trees for utility more often than for biodiversity or carbon. *Biological Conservation* **261**, 109224 (2021).
17. Locatelli, B. *et al*. Tropical reforestation and climate change: beyond carbon. *Restoration Ecology* **23**, 337–343 (2015).
18. Fargione, J. *et al*. Challenges to the reforestation pipeline in the United States. *Frontiers in Forests and Global Change* **4**, 629198 (2021).
19. Kemppinen, K. M. S. *et al*. Global reforestation and biodiversity conservation. *Conservation Biology* **34**, 1221–1228 (2020).
20. Le, H. D., Smith, C., Herbohn, J. & Harrison, S. More than just trees: assessing reforestation success in tropical developing countries. *Journal of Rural Studies* **28**, 5–19 (2012).
21. Hansen, M. C. *et al*. High-resolution global maps of 21st-century forest cover change. *science* **342**, 850–853 (2013).
22. Du, Z. *et al*. Mapping annual global forest gain from 1983 to 2021 with landsat imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2023).
23. Potapov, P. *et al*. The global 2000-2020 land cover and land use change dataset derived from the Landsat archive: first results. *Frontiers in Remote Sensing* **3**, 856903 (2022).
24. Brandt, J., Ertel, J., Spore, J. & Stolle, F. Wall-to-wall mapping of tree extent in the tropics with Sentinel-1 and Sentinel-2. *Remote Sensing of Environment* **292**, 113574 (2023).
25. Brandt, M. *et al*. An unexpectedly large count of trees in the West African Sahara and Sahel. *Nature* **587**, 78–82 (2020).
26. Tucker, C. *et al*. Sub-continental-scale carbon stocks of individual trees in African drylands. *Nature* **615**, 80–86 (2023).
27. Yao, L., Liu, T., Qin, J., Lu, N. & Zhou, C. Tree counting with high spatial-resolution satellite imagery based on deep neural networks. *Ecological Indicators* **125**, 107591 (2021).
28. Velasquez-Camacho, L., Etxegarai, M. & de-Miguel, S. Implementing Deep Learning algorithms for urban tree detection and geolocation with high-resolution aerial, satellite, and ground-level images. *Computers, Environment and Urban Systems* **105**, 102025 (2023).
29. Aziz, G., Minallah, N., Saeed, A., Frnda, J. & Khan, W. Remote sensing based forest cover classification using machine learning. *Scientific Reports* **14**, 69 (2024).

30. Dixit, J. *et al*. Potential of Lightweight Drones and Object-Oriented Image Segmentation in Forest Plantation Assessment. *Remote Sensing* **16**, 1554 (2024).
31. Ecke, S. *et al*. UAV-based forest health monitoring: A systematic review. *Remote Sensing* **14**, 3205 (2022).
32. Hijmans, R., Garcia, N. & Wieczorek, J. GADM: database of global administrative areas, version 3.6. *GADM Maps and Data* (2018).
33. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. CoRR, **abs/1810.04805**, http://arxiv.org/abs/1810.04805 (2018).
34. Sanh, V., Debut, L., Chaumond, J., Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *NeurIPS EMC^2 Workshop* (2019).
35. Pesaresi, M. & Politis, P. GHS-BUILT-C R2023A-GHS Settlement Characteristics, derived from Sentinel2 composite (2018) and other GHS R2023A data. *European Commission, Joint Research Centre (JRC)*. PID: http://data.europa.eu/89h/3c60ddf6-0586-4190-854b-f6aa0edc2a30 (2023).
36. Center for International Earth Science Information Network. Global roads open access data set (gROADS), v1 (1980–2010). *NASA Socioeconomic Data and Applications Center (SEDAC)* (2010).
37. Oogathoo, S., Duchesne, L., Houle, D., Kneeshaw, D. & Nicolas, B. é Precipitation and relative humidity favours tree growth while air temperature and relative humidity respectively drive winter stem shrinkage and expansion. *Frontiers in Forests and Global Change* **7**, 1368590 (2024).
38. Ma, Y., Eziz, A. & Halik, Ü. Precipitation and temperature influence the relationship between stand structural characteristics and aboveground biomass of forests—A meta-analysis. *Forests* **14**, 896 (2023).
39. Harris, I., Osborn, T. J., Jones, P. & Lister, D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data* **7**, 109 (2020).
40. Farr, T. G. *et al*. The shuttle radar topography mission. *Reviews of geophysics***45**, (2007).
41. Cheesman, A. W., Preece, N. D., van Oosterzee, P., Erskine, P. D. & Cernusak, L. A. The role of topography and plant functional traits in determining tropical reforestation success. *Journal of Applied Ecology* **55**, 1029–1039 (2018).
42. Luo, D., Jin, Z., Yu, Y. & Chen, Y. Effects of topography on planted trees in a headwater catchment on the Chinese Loess Plateau. *Forests* **12**, 792 (2021).
43. Gibbs, H. K., Ruesch, A. S. & Achard, Fré Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s. *Proceedings of the National Academy of Sciences* **107**, 16732–16737 (2010).
44. Tucker, C. J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment* **8**, 127–150 (1979).
45. Pansit, N. R. & Parilla, R. B. Detecting Vegetation Cover Change in Reforestation Sites from 2013 to 2019 in Central Visayas, Philippines Using Remotely Sensed Data. *Mindanao Journal of Science and Technology***22** (2024).
46. Kirbizhekova, I. I., Chimitdorzhiev, T. N., Dmitriev, A. V. & Baltukhaev, A. K. Monitoring of post-fire reforestation based on vegetation indices of optical and radio bands. *29th International Symposium on Atmospheric and Ocean Optics: Atmospheric Physics* **12780**, 1293–1296 (2023).
47. Huang, C., Yang, Q. & Huang, W. Analysis of the spatial and temporal changes of NDVI and its driving factors in the Wei and Jing River Basins. *International Journal of Environmental Research and Public Health* **18**, 11863 (2021).
48. John, A. *et al*. A global dataset of location data integrity-assessed reforestation efforts. Harvard Dataverse. https://doi.org/10.7910/DVN/ZJODGO (2025).
49. Carbon Offset Guide. Carbon Crediting Programs. https://offsetguide.org/understanding-carbon-offsets/carbon-offset-programs/.
50. Pokorny, B. Forests as a Global Commons: International Governance and the Role of Germany. *Deutsche Nationalbibliothek* (2019).
51. Simkins, T., Bubb, I., Roberts, K. & Huanca-Nunez, N. Haiti Agriculture: Utilizing NASA Earth Observations to Evaluate the Success of Reforestation Practices in Haiti. *DEVELOP Spring 2022* (2022).
52. Mahato, A. Comparative Assessment and Monitoring Changes in NDVI of Achanakmar Tiger Reserve (ATR) and its Buffer Zone, India. *Nature Environment and Pollution Technology* **22**, 913–919 (2023).
53. Mulla, D. J. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering* **114**, 358–371 (2013).
54. Pettorelli, N., Pelletier, F., Hardenberg, A. V., Festa-Bianchet, M. & Côté, S. D. Early onset of vegetation growth vs. rapid green-up: Impacts on juvenile mountain ungulates. *Ecology* **88**, 381–390 (2007).
55. Cui, P. *et al*. Comparison and Assessment of Different Land Cover Datasets on the Cropland in Northeast China. *Remote Sensing* **15**, 5134 (2023).
56. Venter, Z. S., Barton, D. N., Chakraborty, T., Simensen, T. & Singh, G. Global 10 m land use land cover datasets: A comparison of dynamic world, world cover and esri land cover. *Remote Sensing* **14**, 4101 (2022).

Acknowledgements

The work is supported by funding from the Alexander von Humboldt Foundation and the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) of Germany.

Author contributions

A.J., S.A., T.K., A.T. and I.W. conceived the presented idea and developed the methodology. A.J., S.A. and T.K. acquired the data. A.J. performed the computations. A.J., S.A., T.K., A.T. and I.W. analyzed the results. A.J., S.A. and T.K. wrote the manuscript with support of A.T. and I.W. helped supervise the project. All authors reviewed and edited the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025

Tables

Name	Status	Reason	Collection mode
8 Billion Trees https://8billiontrees.com/planting-projects/	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Afr100 https://afr100.org/	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
All4Trees https://projects.all4trees.org	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
African Development Bank https://mapafrica.afdb.org/en/projects	Unsuccessful	Could not computationally retrieve data	
Americanforests https://www.americanforests.org/	Successful	Point Geometry Available	API
Anthesis https://www.climateneutralgroup.com/en/climate-projects/	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Arbor Day Foundation https://www.arborday.org/	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Atlas https://atlas.openforestprotocol.org/	Successful	Polygon Geometry Available	API
Bôndy https://www.bondy.earth/en/projet	Unsuccesful	Location mentioned but no geometry available	
Climate Action Reserve https://www.climateactionreserve.org/registry/	Unsuccessful	Could not computationally retrieve data	
Climate Impact Partner https://www.climateimpact.com/	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Climate Partner https://climatepartnerimpact.com/projects/	Successful	Polygon Geometry Available	
Cnaught https://www.cnaught.com/projects	Unsuccessful	Could not computationally retrieve data	
Coeur de Foret https://www.coeurdeforet.com/projets	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Ecologi https://ecologi.com/projects	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Ecotree https://ecotree.green/	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Ecosia https://blog.ecosia.org/tag/where-does-ecosia-plant-trees/	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Eden People + Planet https://www.edenprojects.org	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
ExplorerLand https://explorer.land/x/projects	Successful	Polygon Geometry Available	API
Face the Future https://facethefuture.com	Successful	Polygon Geometry Available	API
First Climate https://www.firstclimate.com/klimaschutzprojekte?lang=en	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Forliance https://forliance.com	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Forest-trends https://www.forest-trends.org/project-list/#close	Successful	Point Geometry Available	API
Fund Forest (Conservation) https://www.fundforests.org	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Gold Standard https://registry.goldstandard.org/projects?q=&page=1	Unsuccessful	Could not computationally retrieve data	
Green Climate Fund https://www.greenclimate.fund/projects	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Greenforestfund https://www.greenforestfund.de/en/sites/	Successful	Point Geometry Available	Manually Collected
GrowMyTree https://growmytree.com/en/pages/transparenz#partnerships	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Humy https://www.humy.org/nos-projets	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
IDRECCO https://www.reddprojectsdatabase.org/browse-redd-data/	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Mastreforestation https://www.mastreforest.com/reforestation	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Mongabay https://reforestation.app/explore?search=%22%22&sort=%22context%22	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
mossy.earth https://www.mossy.earth/projects	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
One Tree Planted https://onetreeplanted.org/	Successful	Could not computationally retrieve data	
On a mission https://www.onamission.world/projects	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Plant for Planet https://www.plant-for-the-planet.org/	Successful	Polygon Geometry Available	API
Primaklima https://www.primaklima.org/was-wir-tun	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Reforest Action https://www.reforestaction.com/en/projects?category=FOLLOW_UP	Successful	Point Geometry Available	API
Reforestum https://reforestum.com/	Successful	Point Geometry Available	API
Replant Canada https://www.replant-environmental.ca/nationalparks.html	Unsuccessful	Could not computationally retrieve data	
Terre du Futur https://www.terre-du-futur.fr/projet-de-reboisement-en-france/	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
TIST https://program.tist.org/	Successful	Polygon Geometry Available	
Tree Nation https://tree-nation.com/	Successful	Polygon Geometry Available	API
Trees for Life https://treesforlife.org.uk/support/plant-a-tree/	Unsuccessful	Location Available, Exact Site Geometry Unavailable	
Trees for the Future https://trees.org/about-us/	Successful	Point Geometry Available	API
Treedom.net https://www.treedom.net/	Unsuccessful	Could not computationally retrieve data	
United Nations Offset Program https://offset.climateneutralnow.org/reforestation-and-afforestation	Unsuccessful	Reforestation Data Unavailable	
Verra https://registry.verra.org	Successful	Polygon Geometry Available	Download
Verritree https://www.veritree.com/	Successful	Point Geometry Available	API
Winrock https://acr2.apx.com/myModule/rpt/myrpt.asp?r=111	Unsuccessful	Could not computationally retrieve data	
Zeroco2 https://zeroco2.eco/en/projects/	Successful	Point Geometry Available	API

Table 7. List of websites considered.

Indicator name	Definition	Data Source
planting_date_reported	Date when most of the tree planting took place	Organization metadata
trees_planted_reported	Number of trees planted	Organization metadata
site_sqkm	Size of the site in square kilometres	Organization metadata / GeoPandas
survival_rate_reported	Share of trees survived (tbd years after planting)	Organization metadata
species_planted_reported	Tree species planted at each site	Organization metadata
description_reported	Site description provided by project developer	Organization metadata
project_id_reported	Project ID from project website	Organization metadata
site_id_reported	Site ID from project website	Organization metadata
species_count_reported	Number of species reported planted	Organization metadata
geometry_reported	Geo-coordinates provided by project website	Organization metadata
country	Country of site location	Organization metadata / GeoPandas
url	URL of project organization's page	Organization website
host_name	Name of organization hosting site info	Organization website
Nested_in	IDs of intersecting polygons	—
exact_admin_area	Whether polygon aligns with admin area	GADM
Polygon_acircle_oval_98	Whether polygon is (almost) circular/oval	—
project_geometries_invalid	Whether geometry is valid (polygon/point)	—
species_planted_derived	Species extracted via LLM from PDFs/descriptions	Project PDFs / LLM output
planting_dates_extracted	Planting dates extracted via LLM	Project PDFs / LLM output

Table 8. Indicators sourced from project websites and derived from PDFs.

Indicator name	Definition	Data Source
total_road_length_km	Length of roads within the site (km)	NASA gROADS (Earth Engine)
built_area	Built-up area within the site (sq km)	GHS BUILT Settlement (Earth Engine)
loss_pre_5	Average tree loss five years before planting	GFW Global Forest Change (Earth Engine)
loss_post_3	Average tree loss three years after planting	GFW Global Forest Change (Earth Engine)
loss_post_5	Average tree loss five years after planting	GFW Global Forest Change (Earth Engine)
cropland_from_tree	Forest-to-cropland conversion area (2000-2020)	GLAD (Earth Engine)
cropland_to_tree	Cropland-to-forest conversion area (2000-2020)	GLAD (Earth Engine)
permanent_water	Water body area within the site (sq km)	GLAD (Earth Engine)
short_vegetation_after_tree_loss	Re-vegetation after tree loss (sq km, 2000-2020)	GLAD (Earth Engine)
top_three_ndvi_months	Top three months with highest NDVI in 2023	Sentinel-2 (Earth Engine)
tree_cover_area_2020	Tree cover area in 2020 (sq km)	GLAD (Earth Engine)
tree_cover_area_2015	Tree cover area in 2015 (sq km)	GLAD (Earth Engine)
tree_cover_area_2010	Tree cover area in 2010 (sq km)	GLAD (Earth Engine)
tree_cover_area_2005	Tree cover area in 2005 (sq km)	GLAD (Earth Engine)
tree_cover_area_2000	Tree cover area in 2000 (sq km)	GLAD (Earth Engine)
average_precipitation	Average monthly rainfall at centroid (planting, 1, 2, 5 yr)	WorldClim[39]
tmax_and_tmin	Max/min temperature at planting, 1, 2, 5 yr	WorldClim[39]
Change_1year	Tree-cover change (NDVI + Shadow Index) 1 yr post-planting	COPERNICUS/S2_HARMONIZED
Change_2years	Tree-cover change (NDVI + Shadow Index) 2 yr post-planting	COPERNICUS/S2_HARMONIZED
Change_5years	Tree-cover change (NDVI + Shadow Index) 5 yr post-planting	COPERNICUS/S2_HARMONIZED
mean_elevation	Mean elevation of site (m)	NASA SRTM DEM (30 m)
mean_slope	Mean slope of site	NASA SRTM DEM (30 m)

Table 9. Indicators from secondary spatial and environmental datasets.