Biotropica

WILEY

bioTROPICA ASSOCIATION FOR TROPICAL BIOLOGY AND CONSERVATION

ORIGINAL ARTICLE

# Harmonizing Distributed Tree Inventory Datasets Across India Can Fill Critical Gaps in Tropical Ecology

Krishna Anujan[1,2,3] | Neha Mohanbabu[1,4] | Abhishek Gopal[1,5,6] | Akshay Surendra[1,7,8] | Aparna Krishnan[1,9] | Ankitha Jayanth[1,10] | Tanaya Nair[1,11] | Shasank Ongole[1,12] | Mahesh Sankaran[1,9]

[1]India Tree Inventory Network, National Centre for Biological Sciences, TIFR, Bengaluru, India | [2]Smithsonian National Zoo and Conservation Biology Institute, Front Royal, Virginia, USA | [3]Smithsonian Environmental Research Centre, Edgewater, Maryland, USA | [4]Utrecht University, Utrecht, the Netherlands | [5]CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India | [6]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India | [7]Lewis B. Cullman Fellow, New York Botanical Garden, Bronx, New York, USA | [8]School of the Environment, Yale University, New Haven, Connecticut, USA | [9]National Centre for Biological Sciences, TIFR, Bengaluru, India | [10]University of South Bohemia, Ceske Budejovice, Czech Republic | [11]Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK | [12]Plant Ecology, University of Bayreuth, Bayreuth, Germany

**Correspondence:** Krishna Anujan (krishna.anujan@gmail.com)

**Received:** 24 September 2025 | **Revised:** 4 April 2026 | **Accepted:** 23 June 2026

**Associate Editor:** Katerina Sam | **Handling Editor:** Katerina Sam

**Keywords:** data gaps | FAIR data | India | macroecology | tree inventory

## ABSTRACT

Global macroecological analyses of biodiversity and associated functions are geographically biased with poor representation from many unique and threatened regions, hindering truly global understanding, and effective conservation. Not all regional gaps are equal, but may be driven by biases in sampling and/or data accessibility, necessitating problem-oriented, and regionally-tailored strategies to address them. We examine the causes and consequences of biodiversity data gaps using tree inventories from India, a global hotspot for biodiversity and carbon sequestration, as an example. We systematically collated the India Tree Inventory metadataset, INvenTree-meta, combining literature searches and manual data retrieval and assessed data availability, reusability, and potential for synthetic research. INvenTree-meta is the largest meta-dataset of peer-reviewed. publications ($n = 465$) on geolocated plot-based tree inventories of multispecies communities from Indian ecosystems, in aggregate covering three decades, and all of its woody biomes. We show extensive sampling across biomes but poor data accessibility (73.33% of studies representing 83.43% of the sampled area after removing repeats) through descriptive analyses. However, we observe majority authorship from India; 82.8% of corresponding authors were from India and 73.33% of the studies had all authors affiliated with Indian institutions demonstrating potential for regional collaboration to fill this gap. We illustrate the potential of systematically assembled regional metadatasets to identify future sampling priorities and of a harmonized dataset to advance ecological science, promote sound management and aid education. Assessing global data gaps with a focus on equity can accelerate filling knowledge gaps, as well as informing regional management and conservation.

## 1 | Introduction

Distributed and standardized biodiversity datasets are the cornerstone of macroecological research, yet these datasets have geographic gaps (Martin et al. 2012) for most taxa that hinder truly global understanding. Tree inventory information is a foundational biodiversity dataset that forms a crucial baseline to understand ecological communities, their structure and functioning in the face of global change (Anderson-Teixeira et al. 2015; Blundo et al. 2021; Cazzolla Gatti et al. 2022).

While tree inventories may involve different objectives, plot sizes and design, they are typically collected using one of a few standardized methodologies, allowing for harmonization (e.g., Anderson-Teixeira et al. 2018). Despite widespread efforts and databases that invite voluntary contributions (e.g., Blundo et al. 2021), the global and regional distribution of tree inventory datasets is geographically biased, leading to highly variable and weakly constrained predictions from data-poor areas, many of which overlap deforestation and climate change hotspots (Carvalho et al. 2023). Data gaps might be an outcome of either limited sampling in the region or a lack of accessible and standardized datasets, requiring problem-specific solutions.

Continental biases in data pose a barrier to global ecology, both because of ecological and anthropic differences among the global tropics. Firstly, tropical biomes show continental scale differences in species richness, stem density, and carbon stocks, because of differences in earth history and biogeography (Sullivan et al. 2017; Hagen et al. 2021). Secondly, historical and ongoing human influence can shape natural landscapes, influencing patterns, processes, and functioning but its degree varies considerably across continents and biomes (Venter et al. 2016; Riggio et al. 2020). Yet, global analyses using tree inventories draw on sparse and biased data from the tropics, hindering truly "global" ecological knowledge (Stocks et al. 2008; Martin et al. 2012). Global syntheses use publicly available data (e.g., hosted at GBIF or Dryad) and aggregate national scale data (e.g., US Forest Inventory and Analysis, FIA), inadvertently reflecting and propagating biases within these datasets. Temperate biomes are oversampled, while tropical biomes are understudied, and the Neotropics are studied more than the Paleotropics, which have relatively higher human influence (Stocks et al. 2008; Martin et al. 2012). For a robust understanding of tropical ecology in the Anthropocene, research must span variation in both ecological and human dimensions.

Across the tropics, data accessibility hinges not only on sampling constraints but on legacies of scientific inequalities, further hindering collaboration and synthesis. Stemming from colonial legacies, tropical ecology research and publishing has been dominated by researchers from high-income countries, leading to "parachute science" (Nuñez et al. 2021; Miller et al. 2023; Maas et al. 2021). Despite ecological research moving to a big-data and collaborative discipline, its costs and benefits continue to be unevenly distributed (Maas et al. 2021). Researchers based in the Global South experience multistep challenges in scientific publishing—lack of funding, language barriers, inequalities during peer review and inability to pay high open-access publishing costs—as well as transient and transactional partnerships with Global North researchers (Mekonnen et al. 2021; Asase et al. 2022). Rampant academic inequalities could disincentivise data sharing, especially in the absence of clear incentives (De Lima et al. 2022). Since barriers to data access have strong social dimensions, it is important to carefully examine the geographic distribution of the researchers, biomes and publication patterns to understand the community of data owners and assess collaboration needs.

The Indian subcontinent is a pronounced gap in most global datasets and analyses, despite its high ecological and conservation importance. India, currently the most populous country in the world, continues to maintain large forest cover and high species diversity under four biodiversity hotspots (Myers et al. 2000; Mittermeier et al. 2004). Among the global tropics, this region ranks highest in present-day human footprint on ecosystems, on par with highly built areas of industrialized temperate regions (Venter et al. 2016). Despite this unique combination of bioclimatic, biogeographic and human forces in shaping present-day natural landscapes, the subcontinent continues to be excluded from macroecological syntheses, including estimates of tree density, forest age and species richness (e.g., Cazzolla Gatti et al. (2022); Crowther et al. (2015); Poulter et al. (2019)), potentially biasing global ecological inferences. ForestPlots.net, a voluntary, researcher-contributed database that collates and hosts data on forest monitoring across the tropics reports 7044 plots in 63 countries, concentrated in the tropics. This database, used for several global syntheses, however, includes only 8 plots from the Indian subcontinent (as of February 2024). These geographic gaps in global databases and syntheses could be due to a combination of factors including gaps in data collection within the subcontinent (e.g., across biomes or regions of interest for specific studies) or gaps in data accessibility (e.g., in global databases) gaps in data reusability (e.g., data stored in repositories missing data or metadata components). Paradoxically, Indian scholarship in forest ecology is extensive, with a growing number of research programmes and investment, including over 250 long-term monitoring efforts (Tiruvaimozhi et al. 2025). Given its large ecological and human capital, collaborative frameworks from India can set benchmarks for inclusive tropical ecology.

We assessed the causes and consequences of regional data gaps in ecology through a systematic approach using tree inventories from India. We assembled INvenTree-meta, the largest meta-dataset of geolocated plot-based tree inventories from India and assessed (i) sampling patterns, (ii) data reusability, and (iii) authorship patterns in tree community research in India to aid future syntheses. We also use this dataset to (iv) illustrate applications for identifying sampling priorities for future research.

## 2 | Materials and Methods

### 2.1 | Study Area

The Indian subcontinent is one of the oldest regions in tropical Asia (Mani 1974; Briggs 2003). Its complex geo-climatic and biogeographic history has resulted in a rich variety of biomes with high diversity and endemism of flora and fauna (Figure 1). Within the subcontinent, India, 6° 440 and 35° 300 N latitude and 68° 70 and 97° 250 E longitude, comprises the largest landmass, with four global biodiversity hotspots occurring within mainland India (Myers et al. 2000; Mittermeier et al. 2004). It comprises ten biogeographic subdivisions based on phytogeography and five major broad forest types (Champion and Seth 1968; Rodgers and Panwar 1988) (Figure S2). These classification frameworks that use criteria ranging from geological history to conservation importance can have distinct uses. For ease of comparison with global studies, we used Dinerstein et al. (2017), which resulted in classification of the Indian mainland and the Andaman and Nicobar Islands into ten biomes with several ecoregions nested under these (Figure 1, Figure S3).

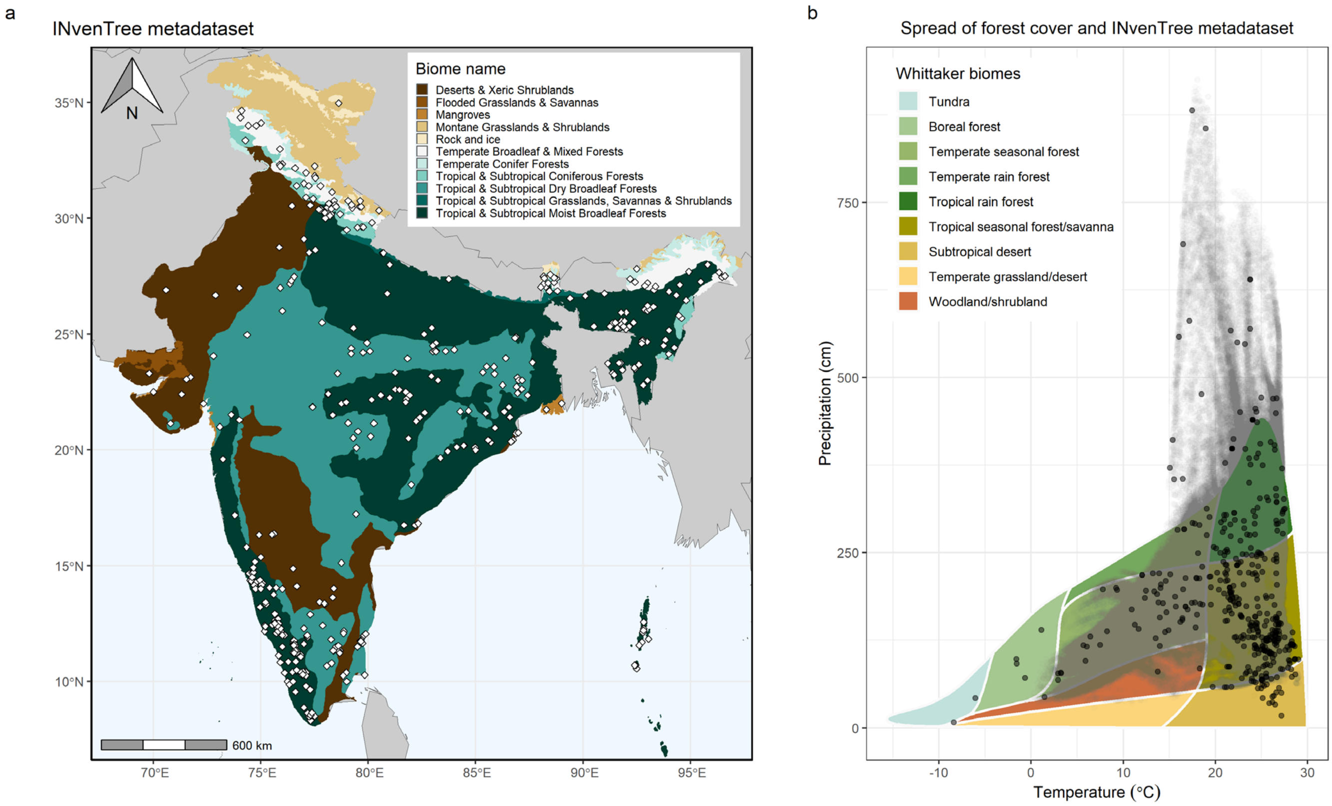


**FIGURE 1** | INvenTree-meta dataset. (a) The location of study sites in papers that form part of the INvenTree-meta dataset in relation to the 10 vegetated biomes (and Rock and Ice) in India as per the Dinerstein et al. (2017) dataset. Each point corresponds to the centroid of the sampled location reported in each study. (b) The distribution of studies (dark gray points) and all forested pixels in India (light gray points) as per Hansen et al. (2013) in the year 2000 with a threshold of > 30% cover per Friedl and Sulla-Menashe (2022) over the distribution of Whittaker biomes. Axes represent mean annual temperature and mean annual precipitation from BioClim.

## 2.2 | Assembling INvenTree-Meta

We systematically assembled a meta-dataset of plot-based tree inventory studies on multispecies communities from Indian forested ecosystems from 1991 to 2023 (Figure S1). We used both Google Scholar (last search in November 2021) and Web of Science (last search in February 2023) to include a wide range of journals as well as book chapters that publish forestry research from India. The Google Scholar search terms included (Tree diversity OR plot OR dbh) AND (India) and excluded other countries where "India" or "Indians" may have been used in other contexts (e.g., other countries in the Indian sub-continent, American Indian reservations). We used a broader list of search terms on Web of Science with the keywords AB = ((tree diversity OR forest structure OR (tree* AND biomass) OR (forest* AND biomass) OR carbon stock OR vegetation survey OR vegetation sampling OR (tree* AND plot) OR (forest AND plot) OR (measur* AND tree) OR (checklist AND flora) OR (checklist AND tree) OR (checklist AND plant*)) AND India). We supplemented this list with an additional set of studies we were aware of but weren't present in the web search (e.g., global studies that report on datasets from India but are not captured by the keywords), and with studies that we found through snowball sampling the known papers' references. We created a merged list ($n = 3758$) removing duplicates and entries that were not relevant based on the title and abstract, resulting in 657 potentially relevant papers. We accessed shortlisted papers and screened them based on the following criteria– (1) the study involved collection of plot-level or checklist data of trees, and (2) the study sites were natural multispecies stands of trees from inland woody ecosystems or coastal forests and mangroves. We excluded papers focussing on seedlings, single-species studies and those based in managed or urban environments. Based on these criteria we identified 465 papers that formed the INvenTree-meta dataset.

From each paper in INvenTree-meta, we extracted data on (i) geographic location of study (ii) area sampled (iii) author affiliations and (iv) data accessibility. (i) We recorded the latitude and longitude of the study area from the text or from a map provided. For studies with multiple plots with separate locations reported, we recorded the study centroid. As this can introduce uncertainty in groupings, studies that span multiple biomes ($n = 5$) or administrative units ($n = 9$) were noted and excluded from specific analyses as needed. (ii) To calculate the total area sampled for plot-based studies, we recorded plot shape, plot dimensions and number of plots. For studies that sampled subplots within a larger plot (e.g., five 20 m x 20 m plots sampled within 1 ha), we recorded the total area of the subplots which were exhaustively sampled ($5 \times 400$ m$^2$) instead of the reported study area (1 ha). For studies that inventoried plots across natural and managed/planted communities, we only counted plots that were in natural multispecies communities. (iii) We verified or manually added information on author affiliations and addresses. (iv) We recorded whether the papers were freely available (open access) on the publisher's webpage and whether at least plot-level data were accessible in the text, appendices, or data repositories like Dryad or Zenodo. Finally, we identified repeat uses of the same datasets—either long-term datasets or syntheses. For repeat datasets we included only the first (earliest) published study in all summaries. For analyses of authorship, we used all published papers, even if the same dataset was used in multiple publications or if multiple datasets were combined, into large-scale and synthetic papers.

### 2.3 | Analyzing Data Accessibility and Reusability

We first examined the number of studies, sampled area (total and percent), article and data access status of studies across biomes. Although some of our tree inventory sites fell outside of biomes expected to have tree cover (e.g., in Montane grasslands and shrublands), we included them as this information is likely to be more accurate than biome classifications. To examine practical constraints on data reuse, we assessed accessible datasets on two axes: data resolution or "graininess" and reusability. We considered these axes independently because for full reusability for synthesis, data should be available at the finest grain, stem level data, but some synthesis might be possible with aggregate data as well. We used the FAIR framework for scientific data management (Wilkinson et al. 2016) to assess the reusability status of datasets in INvenTree-meta, by categorizing datasets in each study based on whether they are Findable, Accessible, Interoperable and Reusable. We considered three data resolutions: plot-level (data aggregated at the plot level, where plot is the smallest spatial unit of sampling), species-level (data as data aggregated at the species level, whether within or across plots), and stem-level data (data reported at the level of data collection—individual stems). For detailed methods, refer to the Appendix S4.

### 2.4 | Analyzing Authorship Patterns

We considered the full INvenTree-meta ($n=465$) for the bibliometric analysis. For each unique study, we extracted the addresses of all authors and the corresponding author. Some publications in our dataset did not have an identifiable corresponding author, in which case it was left empty. In cases where the corresponding author had multiple affiliations, we considered the first. We ran all unique addresses from papers through a Google Maps-based geocoding software, Awesome Table Geocode, to extract latitude and longitude (https://awesome-table.com).

To analyze collaboration potential within the region for future syntheses, we compared the geographic positions of the authors and study sites using studies where we had information on both ($n=456$). First, we examined the proportion of authors in each study that were affiliated with institutions in India and its relation to the area sampled. Second, we examined the distance of the study site from the corresponding author institution for each study, by calculating the Haversine distance (the shortest distance between the two points assuming a spherical earth) between the two pairs of coordinates using the *geosphere* package (Hijmans 2022a), and considered its untransformed distribution. While true authorship, leadership, and data ownership dynamics cannot be captured by this simplified metric, it can provide a coarse approximation of intellectual ownership and is indicative of parachute science.

### 2.5 | Applications: Analyzing Geographic Patterns in Sampling

We used INvenTree-meta to illustrate a potential heuristic for sampling priorities in tree inventory research in India. We assessed sampling priorities in India in three different ways; we compared the spread of INvenTree-meta in relation to (a) tree cover (as a proxy of ecological sampling priority) and (b) its loss in India (conservation sampling priority) and we identified overlaps in these categories to describe (c) combined sampling priorities. Indian forested ecosystems have a large range of natural tree cover, spanning closed canopy forests to open natural ecosystems. We defined high ecological priority sampling areas as undersampled areas with high tree cover and high conservation priority sampling areas as areas of tree cover loss but no sampling effort. We also defined combined high priority sampling areas as areas that were defined as high priority under both classifications. We used the Hansen et al. (2013) dataset of tree cover in the year 2000 (within the period of interest) and tree cover loss from 2000 to 2021 cropped to the limits of India using Google Earth Engine (code here—https://code.earthengine.google.com/1f887c3af348300241dfc60e3da21cff). For these 30 m × 30 m pixels, we followed Friedl and Sulla-Menashe (2022)'s definition of forested pixels as > 30% forest cover. This would include woody savannas (30%–60% cover) but exclude more open and grassy savannas (10%–30%).

India has 594 districts as of the 2011 census with a median area of 4077.48 $km^2$ ranging from 15.7 to 47,858 $km^2$ area. We chose district as our sampling unit as the sampling scale of most of the studies in INvenTree-meta was contained within a district. Moreover, districts span limited ecological heterogeneity and summaries at this scale are meaningful for further use by multiple stakeholders (e.g., Srivathsa et al. 2023). We used an administrative map of districts accessed from the Database of Global Administrative Areas (GADM) and extracted the tree cover (number of forested pixels) and tree cover loss (number of pixels experiencing loss) for each district using the *extract* function from the *terra* package (Hijmans 2022b). For each district, we then calculated three variables—percent tree cover (percentage of district area with tree cover), percent forest loss (number of pixels that experienced forest loss divided by number of forested pixels) and proportion of sampled area. We excluded large-scale datasets from analysis of biome-distributions as they covered several biomes; for long-term datasets we included only the first (earliest) published study. We limited the analysis to districts that have nonzero forest cover ($n=341$) and excluded all points that fell outside these even though they may have reported on tree-based biomes, because of high uncertainties.

We assigned each district into one of two categories along each variable assessed—% tree cover, % forest loss and sampling effort. While continuous data can provide more information, categorizing can provide clearer paths to decision-making and prioritization that rely on relative rankings. We chose two categories to provide a simple and understandable visualization scale while capturing the largest differences in forest cover, loss and sampling effort. For percent tree cover, we split districts into two even quantiles at the median (12.94%), that is, districts that fell in the lower half of % tree cover values (median = 1.64%) were scored as 0, and the higher half (median = 52.91%) was scored as 1. For tree cover loss, we categorized districts with zero stand-clearing loss as 0 ($n = 115$), and the remaining districts with nonzero loss (median loss = 4.7%) as 1. Since we were interested in defining sampling gaps, districts with zero sampling effort were categorized as 1 (=sampling gap, $n = 209$), and the remaining districts with nonzero sampling were categorized as 0. In our scale, a district with high tree cover and no sampling effort would be scored as high priority ecological sampling. Similarly, a district with nonzero loss and no sampling effort would be scored as high priority conservation sampling. Finally, a district classified as both high priority ecological sampling and high priority conservation sampling was categorized as combined high priority sampling. We visualized spatial overlaps in tree cover or tree cover loss and sampling gaps using bivariate chloropleth maps. Since there are several ways to categorize these continuous variables, we also provide an alternate visualization using $3 \times 3$ categories in the Appendix S7: Figure S16.

All analyses were performed in R version 4.1.3 (R Core Team 2022) using packages *dplyr* and *tidyverse* (Wickham et al. 2019, 2022).

## 3 | Results

### 3.1 | Data Availability and Accessibility

The 465 studies reported in the journal articles in INvenTree-meta span 10 biomes and cover 31 of the 36 states and union territories of India (Figure 1). We identified 436 studies of this dataset that report on area-based inventories, for example, transects or plot-based measures, sampling a total of 4653.64 ha (Figure 1). The remaining were checklists, duplicates, or did not report key metrics to estimate sampling effort. The most commonly studied biome was tropical & subtropical moist broadleaf forests (236 out of 465, 50.75%), followed by tropical & subtropical dry broadleaf forests (86 out of 465, 18.49%) (Figure 2a). In terms of area, the most extensively sampled biome was tropical & subtropical moist broadleaf forests with a total area of 2169.3 ha sampled across 221 studies (Figure 2b).

Most studies in INvenTree-meta sampled limited area; the median sampled area of the studies was 2 ha. Of the total studies, 84.73% report on datasets that sample less than a total area of 100 ha, 72.47% < 10 ha, 27.96% < 1 ha. Across these studies, data accessibility was poor; only 101 studies, representing 21.72% of the studies and 14.8% of the area had data that was available either as a table, appendix, supporting information or archived in a repository. Sampled area was not statistically significantly different between data-accessible and data-inaccessible studies (Kruskal–Wallis test, chi-squared = 1.3117, df = 2, $p = 0.519$). We report a nonparametric test because while the groups meet homogeneity assumptions (Levene's test, $F$-statistic = 0.29, df = 2, $p = 0.74$) the data are not normally distributed (Shapiro–Wilk test, $W = 0.153$, $p < 0.001$). Data access was fairly evenly limited across the biomes studied ($\chi^2$ test statistic for number of studies = 40, $p = 0.1$, for area = 80, $p = 0.2$) (Figure 2b).

We found considerable variation in the level of detail of findable and reusable data across studies that had "accessible" data (Figure 2c). Many "accessible" datasets (14.8% of the total sampled area) met FAIR reusability standards only for plot-level (12.81% of total sampled area across studies) or species-level data (1.95% of sampled area). Among the studies assessed, only five representing 0.47% of sampled area met reusability standards for stem-level data, the finest grain of tree inventory information assessed. Despite limited data accessibility, the text of 56.77% of published articles was openly accessible (Figure S5). Studies were most commonly published in the journal Tropical Ecology ($n = 40$, Figure S5).

### 3.2 | Authorship Patterns

Over 4 in 5 of the corresponding authors of papers in INvenTree-meta were based in institutions in India (82.8%), with a small proportion of corresponding authors (14.19%) from other countries, primarily the United States (4.73%), France (2.37%), and China (1.72%) (Figure 3). Within India, affiliations were spread widely across the country, showing authorship and potential intellectual and data ownership across institutions (Figure 3a). Publishing rates have increased over the years, much of this attributable to corresponding authors based in India (slope for corresponding authors based in India = 0.819, not based in India = 0.099) (Figure S6). A clear majority, 1612 out of a total 1870 authors across articles in the meta-dataset were affiliated with Indian institutions and organizations ('domestic' in Figure 3), but 24.95% of papers also had authors affiliated with non-Indian institutions ('foreign' in Figure 3). When scored for the percentage of Indian-affiliated authors in a study, 90.11% of studies had more than 50% authors affiliated to India while 73.33% studies had all authors affiliated to Indian institutions. Despite few large-scale studies, they often had high or near-total in-country authorship (Figure 3). Large-scale studies increased over the years (Figure S6), but local and small-scale studies also continue to be published (Figure 3). Further, most studies in INvenTree-meta were conducted close to the institution of the corresponding author; the median distance between corresponding author affiliation and study site was 280.38 km with interquartile range 861.3 km (Figure 3).

### 3.3 | Sampling Priorities

Using INvenTree-meta, we present a heuristic for determining sampling priority regions for ecological and conservation research (Figure 4). INvenTree-meta showed disproportionate sampling across districts, with most (209 out of 341) districts with tree cover never sampled (Figure S12). Proportion of tree cover across districts was highly skewed, with 50% of the districts having less than 13% tree cover. Most districts ($n = 226$) experienced at least 1 $km^2$ of stand-clearing loss from 2000 to 2021, and many districts ($n = 86$) experienced greater than 10% of forest area loss. When sampling gaps were visualized along categories of tree cover and tree cover loss, high heterogeneity

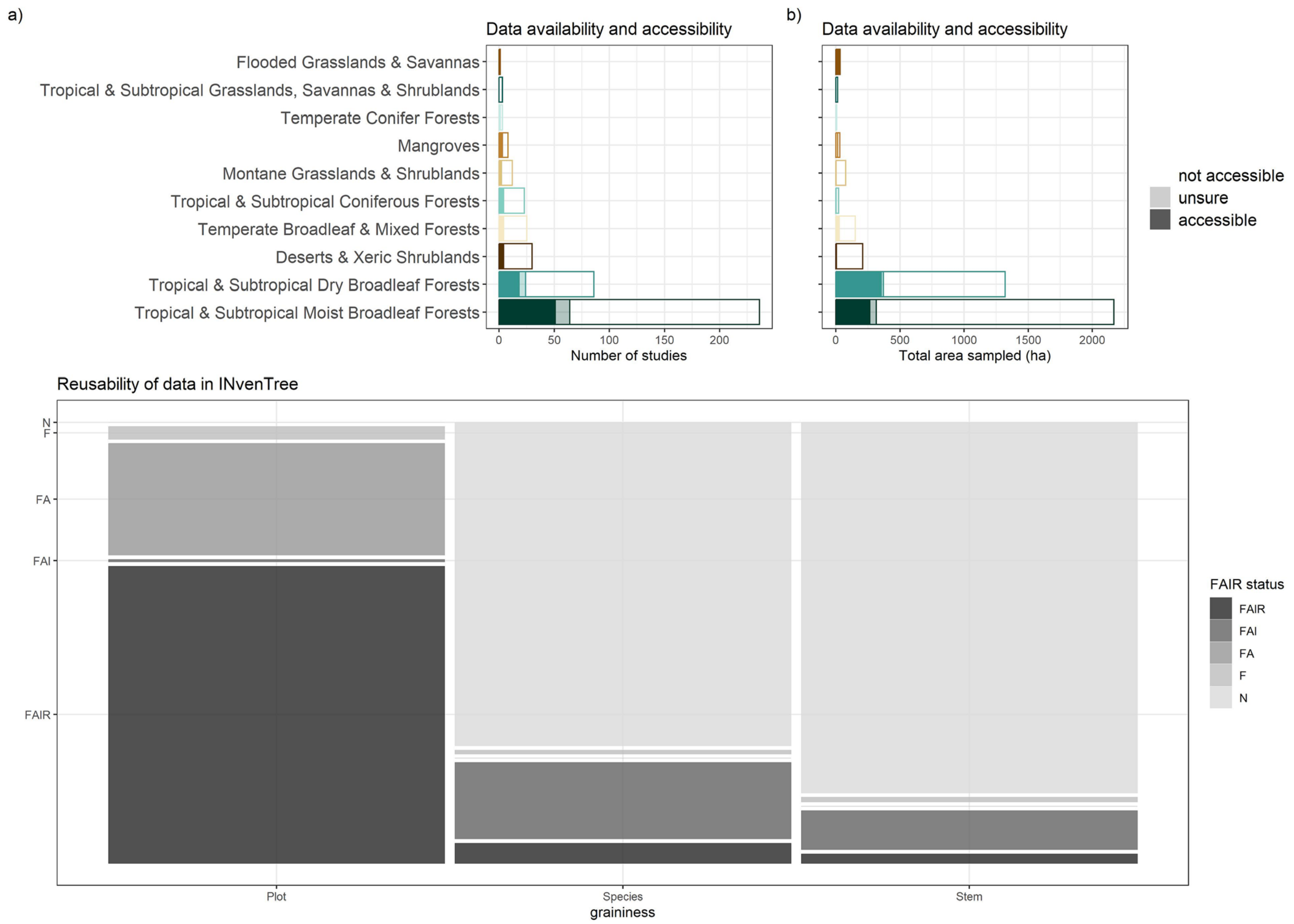


**FIGURE 2** | Data availability and accessibility across biomes. (a) Number of studies and (b) total area sampled across the 10 vegetated biomes of India among tree inventories studies in India. For panels (a) and (b), solid bars represent data accessible in repositories or in the appendix while empty bars represent inaccessible data. Lightly colored bars represent studies where the accessibility status was uncertain. Only studies focussing on a single biome are reported here. (c) Data distribution along data graininess (see methods for more details) and Findable, Accessible, Interoperable, Reusable (FAIR) principles of data reuse, showing low reusability. Data reported only for studies where data was accessible for at least one of the categories of data graininess, the rest has been excluded from this figure. Legends F, FA, FAI, FAIR represent position along the reusability scale as defined by these principles, with corresponding colors from light to dark gray. N represents data that were not findable. Size of the bars represent total sampled area in each of these categories. For detailed methods, refer to the Supporting Information.

among sites and biomes emerged (Figure 4). High priority regions for ecological sampling priority were concentrated in the northern Western Ghats, northeast India, the Nicobar Islands, parts of the Western Himalayas, and a few districts in eastern India that have high forest cover but were never sampled in INvenTree-meta (Figure 4, Figure S12). A few datasets reporting on tree-based inventories fell outside of these forested districts (Figure S7). Sampling effort across ecoregions was highly skewed, and most regions were highly undersampled relative to their extent (Figure S17).

Districts in the highest quantile of forest loss relative to forest cover were clustered mainly in north-east India and the eastern states of Odisha, Andhra Pradesh, Telengana, and Chattisgarh (Figure S14). Among districts that experienced forest loss, districts in north-east and eastern India and the northern Western Ghats were the least studied (Figure 4). When examined together, 79 out of 341 forested districts were identified as high priority sampling for both ecology and conservation. These districts were concentrated in north-east India, central Western Ghats, and eastern India, along with a few districts in the western Himalayas as well as the Nicobar Islands.

## 4 | Discussion

Through a systematic approach, we generated INvenTree-meta, which to our knowledge, is the first and largest meta-dataset of tree inventory information from India. INvenTree-meta spans all tree-based biomes of India and a total of 4653.64 ha, but up to 80% of the studies in the dataset did not have accessible data. Even when accessible, only a strikingly small proportion met reusability standards for any level of data; only five studies representing 0.47% of the total sampled area across studies had data that met FAIR reusability standards for stem-level data, the finest grain of tree inventory information assessed. While reasons for the lack of inclusion of the region in global analyses were not assessed, this result alludes to poor data accessibility as a potential major deterrent. However, most studies in INvenTree-meta are authored by researchers affiliated with Indian universities

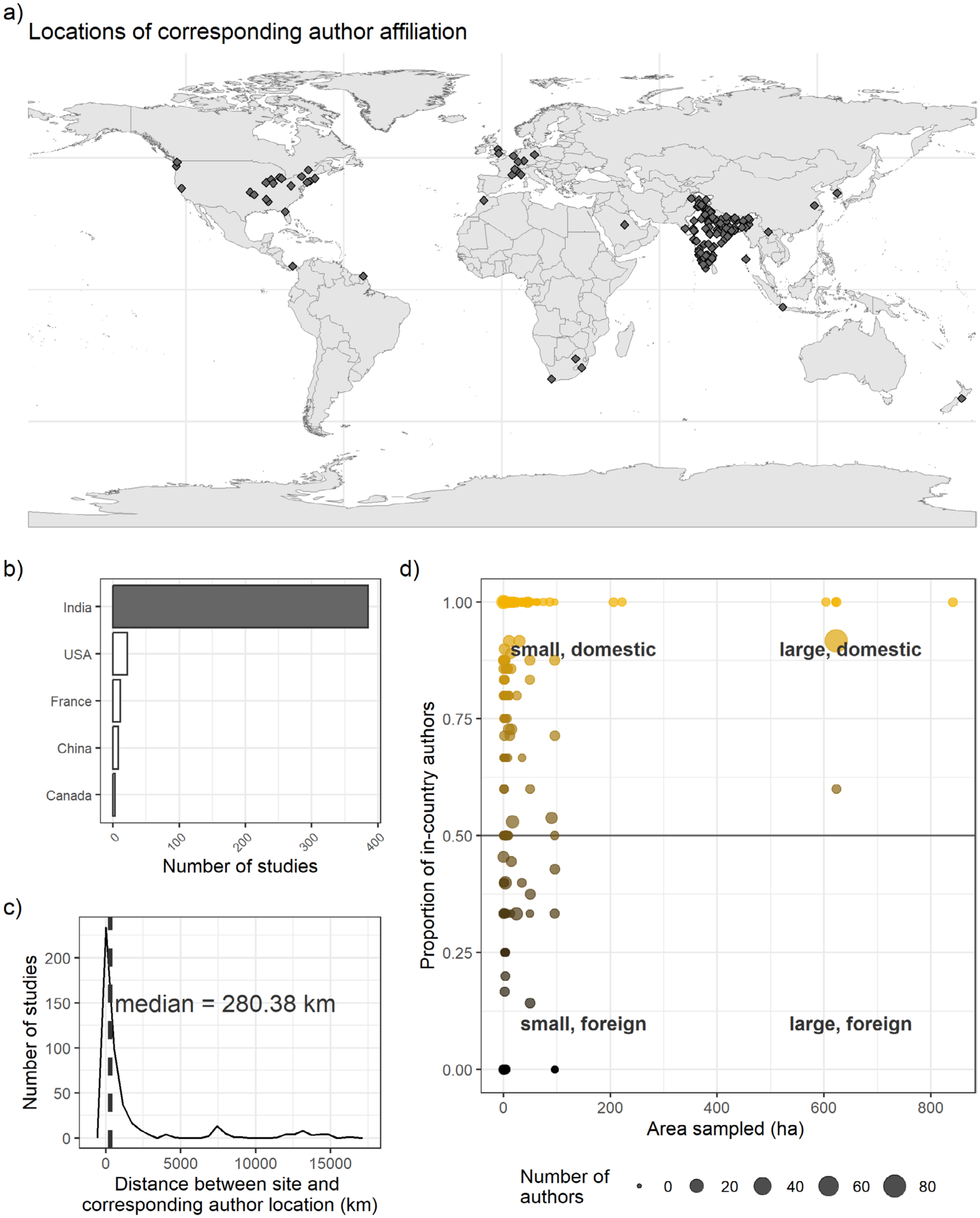


**FIGURE 3** | INvenTree-meta authorship patterns and data equity. (a) Locations of corresponding authors of papers in INvenTree-meta, (b) top five countries, representing 92.26%, where corresponding authors were affiliated; (c) Distance between corresponding author affiliation and study site across studies. (d) Proportion of domestic and foreign authors in INvenTree-meta and their relationship to area sampled. Size of the points represents number of authors in the study, color is indicative of the proportion of in-country authors.

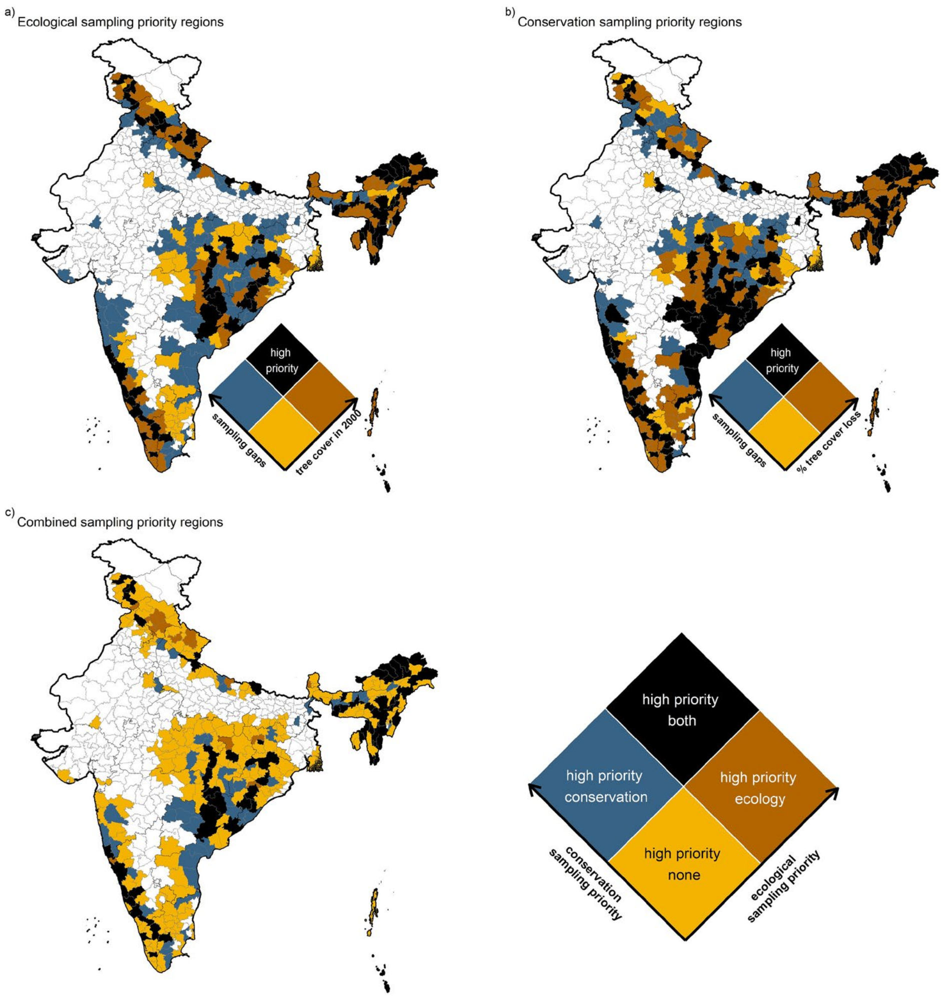


**FIGURE 4** | Ecological, conservation and combined sampling priority regions. Bivariate chloropleth maps showing sampling gaps in plot-based tree inventories across districts in India in relation to (a) mean tree cover in the district in 2000 and (b) mean tree cover loss from 2000 to 2021 averaged over the area of the district. (c) Combined map showing overlap in ecological and conservation priority districts. This map is intended as a heuristic and to illustrate uses of the INvenTree metadataset. Here, ecological sampling priority indicates undersampled areas with respect to tree cover, while conservation sampling priority areas are sites with tree cover loss but no sampling effort. Black color indicates areas with high priority sampling, that is, no sampling and (a) high % forest cover or (b) nonzero forest loss or (c) high % forest cover and forest loss. Yellow indicates low priority sampling regions with nonzero sampling in INvenTree meta and (a) low forest cover or (b) no reported forest loss or (c) either low forest cover or no loss. Brown and darker blue indicate medium priority regions, while districts in white were excluded from the analysis as they did not include forested pixels. Refer methods and supplementary for details.

and study ecosystems close to their home institutions. The strong presence of data owners in Indian organizations and high-quality local data indicates potential for a regional network to promote a better understanding of ecological processes across larger spatial scales. We illustrate potential applications of the metadataset and avenues for regional collaboration by using INvenTree-meta to create a heuristic to identify geographic gaps in sampling and define ecological and conservation sampling priority regions. Harmonizing these datasets under fair sharing protocols that also benefit researchers in the global South is a first step towards synthesis research in the region. Given India's high biodiversity and human-dominated landscapes, such a network could promote a better understanding of ecosystems in the Anthropocene, while benefiting research and researchers.

Using INvenTree-meta, we identified sampling extent and its biases across biomes and regions, providing directions to align future sampling of tree-based biomes within the country (Figure 4). INvenTree-meta spans all of India's diverse tree-based biomes, but shows nonuniform sampling across these, with moist and dry forests most well-sampled relative to their area. Despite their sampling extent, the distribution of plots in moist forests was clustered in the more speciose areas—southern Western Ghats and north eastern forests (Figure 1). Consequently, our approach using INvenTree-meta proposes ecological sampling priority in the northern Western Ghats and north-east India and conservation sampling priority in north-east India as well as eastern India (Figure 4). Regional disparities in sampling can lead to biased understanding of ecological heterogeneity in patterns and processes which may result in inappropriate management, conservation or restoration practices. It is concerning that our proposed regions of priority conservation sampling (high loss/low sampling) overlap regions of large conservation concern with megaprojects (e.g., dams and oil palm expansion in North-East India, Bhattacharya (2022), Srinivasan et al. (2021); proposed development in the Nicobar Islands, Haque and Nawaz (2025); mineral resource extraction in eastern India, Kapoor et al. (2025)). As these regions continue to lose tree cover, there is an urgent need to collect baseline information to support stakeholders in appropriate conservation action. Limitations to this analysis include limitations of the metadataset—centroids were used to represent the spatial extent of the study area and the metadataset does not include gray literature, unpublished data from plots, theses etc.—and limitations of analysis choices—we used a global dataset of tree cover and loss (Hansen et al. 2013) that may have regional errors, and we used a coarse categorical scale to define sampling priorities. However, our proposed ecological, conservation and combined sampling priority maps can provide a blueprint to steer the next decades of tree sampling in the country to help refine our ecological understanding and inform management and conservation.

### 4.1 | Challenges and Promises of Synthesizing INvenTree Data

Large-scale datasets are more than the sum of their parts; for example, data published by Ramesh et al. (2010), spanning large scales in the Western Ghats, have spurred several regional and global syntheses ranging from patterns and drivers of diversity to drivers of above ground carbon storage (e.g., Hardy et al. 2012; Osuri et al. 2020; Davidar et al. 2018). Spatially distributed datasets with standardized sampling like the Forest Inventory and Analysis (FIA) in the United States have led to at least 180 publications and improved predictive accuracy for ecological communities (Tinkham et al. 2018). Similarly, the Forest Survey of India (FSI) conducts countrywide tree inventories spanning ~3500 ha total area every 5 years with a systematic gridded approach (https://fsi.nic.in/about-forest-inventory; Tewari 2016). On the other hand, studies aggregated in INvenTree-meta implement diverse sampling protocols and represent nonuniform and opportunistic spatial spread. However, studies in INvenTree-meta in aggregate cover more than this area even after accounting for sites sampled multiple times. Furthermore, if more datasets are added to this metadata set in the form of contributions from unpublished theses and reports, there is tremendous potential to expand the representation of forests across space and time.

INvenTree-meta highlights the potential for regional or national collaborative approaches to facilitate inclusion of high-quality data into regional and global syntheses. Most research represented in INvenTree-meta was conducted locally, likely integrating deep knowledge of landscape context, taxonomy, and natural history, all crucial for data quality. However, at present, the reusability of these data for further analyses is hindered by the small scale of individual datasets and their distributed ownership; the effort required to identify datasets, data owners, request permissions, and harmonize these may have outweighed the benefits of their inclusion into previous global syntheses. Moreover, since we estimated potential data ownership from corresponding authorship, this may be a poor approximation of true data stewardship, which may be distributed across several authors, groups, and institutions. An accepted set of best practices for data reusability are described by the FAIR principles for data: Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). Our assessment of the reusability of these datasets revealed that even while datasets were "findable" and "accessible", the current forms in which they are publicly available most often did not meet reusability standards, especially for the finest grain, stem-level data (only five studies representing 0.47% of sampled area), required for many syntheses (Figure 2c). This very limited direct reusability of fine-grained data is contrary to current publishing norms in major journals that strongly encourage or mandate data sharing, but the reasons for this pattern were not explored in this study. However, many studies in INvenTree-meta were not only small-scale, but consequently published in regional or specialized journals that may not have such policies in place. Further, those that made data available often made plot or species level data available, which matched the scale of the analyses in the original papers, and perhaps meeting journal requirements. Despite the limited access at the stem level, plot-based vegetation studies are multi-layered in their data as well as further utility, allowing useful, scalable, synthetic research even with plot and species level data. Harmonizing these heterogeneous datasets to achieve "Interoperability" and "Reusability" would require close, continuous collaborations with data owners and collectors to interpret the methodology, consult on taxonomy, and local names of species.

The distributed inventories collated in INvenTree-meta are important not just because of their extent, but the perspective of data sovereignty, and equity. We are encouraged by our findings that showcase a steady pattern of in-country authorship on tree-inventory-based publications from India, a positive contrast to the global pattern of parachute research in tropical landscapes (Asase et al. 2022; Miller et al. 2023). While this pattern, derived from corresponding authorship and affiliations, may not represent true data or intellectual ownership or indicate equal global research partnerships, it highlights the presence of strong domestic scholarship which can be the foundation to establish practices that promote synthesis science and data sovereignty in the region. Based on our findings, we make four recommendations to realize these goals. (1) We propose the India Tree Inventory Network (https://india-inventree.github.io), a regional network of forest researchers from India that will promote synthetic forest research in India and prioritize data sovereignty and knowledge equity at the global and national scale. While recent work in the country has included synthesis on tree diversity (Mugal et al. 2023) and long-term forest studies (Tiruvaimozhi et al. 2025), a formal network can create a collaborative platform, beyond data collection and collation. (2) Costs and benefits of collaboration in the network need to be fairly and equitably distributed among data collectors, owners and users. FAIR principles for data sharing are not always fair for tropical forest data owners and making data unconditionally open could lead to further extractive science, as quantitative skills and publishing power tends to be concentrated in the global north and in elite universities (Mori et al. 2015; De Lima et al. 2022). Existing networks respond to this inequity in diverse ways (e.g., ForestPlots insists co-authorship for primary data collectors (Blundo et al. 2021)), providing models to follow. (3) Collaboration across distributed datasets would require enormous harmonizing efforts, the burden of which cannot fall solely on data producers. Acknowledging the enormous efforts and potential needs of current and future data owners, we envision a regional network that prioritizes quantitative processing tools and support (e.g., open source code, tutorials, periodic data workshops). Combining tools and networks can empower Indian researchers, especially early-career researchers and those in underserved universities to access and utilize these data for research and conservation. (4) We envision a network that actively identifies and supports crosscutting applications of tree inventory data in research, conservation and education by facilitating ease of synthesis, thus filling crucial knowledge gaps in forest ecology. We demonstrate one such application to define sampling priorities (Figure 4), but such data could enhance conservation efforts both by government and private actors by improving data available for monitoring and evaluation of forest management, restoration efforts, carbon credit projects and other c solutions to mitigate climate change. These data would also be useful for universities and educators, enabling data-driven educational products for ecology courses in India (e.g., Organization for Tropical Studies courses in other geographies). We also envision that INvenTree data would serve as crucial baselines for practitioners and local communities. Such continuing and repeated uses of data can feed back to encourage further data collection, sharing and synthesis.

In conclusion, we systematically assess the causes and consequences of regional data gaps in biodiversity baseline datasets using tree inventory datasets from India as an example. Using the first systematic metadataset of tree inventory data from India, we identified the cause of data gaps from this region (limited data reusability) but showcase the potential of a regional network to catalogue and harmonize (meta/)datasets, create and implement future sampling strategies, and expand the scope of macroecological datasets for science, conservation, and education. We also recommend that such networks build on existing frameworks for data and knowledge sharing that acknowledge diverse contributions, costs, and benefits inherent to an unequal world (Cooke et al. 2021; De Lima et al. 2022; Ocampo-Ariza et al. 2023).


## Acknowledgments

We are grateful to the community of data collectors and owners contributing to forest ecology in India. K.A. acknowledges the support of the Smithsonian Institution Postdoctoral Fellowship during the completion of this manuscript. We thank T R Shankar Raman, Divya Mudappa, Anand Osuri, and Navendu Page for their suggestions during the conceptualisation of this project and on the manuscript.


## Conflicts of Interest

The authors declare no conflicts of interest.

## Data Availability Statement

The data and code used in this manuscript are openly available on Zenodo. Data is available with https://doi.org/10.5281/zenodo.15358405 and code is available with https://doi.org/10.5281/zenodo.20827852.

## References


Anderson-Teixeira, K. J., S. J. Davies, A. C. Bennett, et al. 2015. "CTFS-ForestGEO: A Worldwide Network Monitoring Forests in an Era of Global Change." *Global Change Biology* 21: 528–549.

Anderson-Teixeira, K. J., M. M. H. Wang, J. C. McGarvey, et al. 2018. "ForC: A Global Database of Forest Carbon Stocks and Fluxes." *Ecology* 99: 1507.

Asase, A., T. I. Mzumara-Gawa, J. O. Owino, A. T. Peterson, and E. Saupe. 2022. "Replacing "Parachute Science" With "Global Science" in Ecology and Conservation Biology." *Conservation Science and Practice* 4: e517.

Bhattacharya, S. 2022. "India Wants Its Northeast to Be Its "Powerhouse." What of the Indigenous People? – *The Wire Science*."

Blundo, C., J. Carilla, R. Grau, et al. 2021. "Taking the Pulse of Earth's Tropical Forests Using Networks of Highly Distributed Plots." *Biological Conservation* 260: 108849.

Briggs, J. C. 2003. "The Biogeographic and Tectonic History of India." *Journal of Biogeography* 30: 381–388.

Carvalho, R. L., A. F. Resende, J. Barlow, et al. 2023. "Pervasive Gaps in Amazonian Ecological Research." *Current Biology* 33: 3495–3504.e4.

Cazzolla Gatti, R., P. B. Reich, J. G. P. Gamarra, et al. 2022. "The Number of Tree Species on Earth." *Proceedings of the National Academy of Sciences of the United States of America* 119: e2115329119.

Champion, S. H. G., and S. K. Seth. 1968. *A Revised Survey of the Forest Types of India*. Government of India Press.

Cooke, S. J., V. M. Nguyen, N. Young, et al. 2021. "Contemporary Authorship Guidelines Fail to Recognize Diverse Contributions in Conservation Science Research." *Ecological Solutions and Evidence* 2: e12060.

Crowther, T. W., H. B. Glick, K. R. Covey, et al. 2015. "Mapping Tree Density at a Global Scale." *Nature* 525: 201–205.

Davidar, P., F. Munoz, J.-P. Puyravaud, D. Mohandass, and V. S. Ramachandran. 2018. "Multiple Facets of Rarity Among Rain Forest Trees in the Western Ghats of India." *Biological Conservation* 228: 110–119.

De Lima, R. A. F., O. L. Phillips, A. Duque, et al. 2022. "Making Forest Data Fair and Open." *Nature Ecology & Evolution* 6: 656–658.

Dinerstein, E., D. Olson, A. Joshi, et al. 2017. "An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm." *Bioscience* 67: 534–545.

Friedl, M., and D. Sulla-Menashe. 2022. "MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061."

Hagen, O., A. Skeels, R. E. Onstein, W. Jetz, and L. Pellissier. 2021. "Earth History Events Shaped the Evolution of Uneven Biodiversity Across Tropical Moist Forests." *Proceedings of the National Academy of Sciences of the United States of America* 118: e2026347118.

Hansen, M. C., P. V. Potapov, R. Moore, et al. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." *Science* 342: 850–853.

Haque, M., and F. Nawaz. 2025. "A "Development Disaster" in the Making: Focusing on the Mega-Development Project in Andaman and Nicobar Islands, India." In *Disasters, Climate Change and Sustainability in South Asia*, edited by M. Haque, M. Begum, and F. Nawaz, 139–158. Springer Nature Switzerland.

Hardy, O. J., P. Couteron, F. Munoz, B. R. Ramesh, and R. Pélissier. 2012. "Phylogenetic Turnover in Tropical Tree Communities: Impact of Environmental Filtering, Biogeography and Mesoclimatic Niche Conservatism." *Global Ecology and Biogeography* 21: 1007–1016.

Hijmans, R. J. 2022a. "Geosphere: Spherical Trigonometry."

Hijmans, R. J. 2022b. "Terra: Spatial Data Analysis."

Kapoor, S., M. Banerjee, and A. Singh. 2025. "Mining the Mahanadi: The Contested Terrain of Coal Extraction, Ecological Sustainability, and Community Rights in Eastern India." In *Rivers Unbound*, 209–215. Routledge.

Maas, B., R. J. Pakeman, L. Godet, L. Smith, V. Devictor, and R. Primack. 2021. "Women and Global South Strikingly Underrepresented Among Top-Publishing Ecologists." *Conservation Letters* 14: e12797.

Mani, M. S. 1974. *Ecology and Biogeography in India*. Dr W. Junk b.v. Publishers.

Martin, L. J., B. Blossey, and E. Ellis. 2012. "Mapping Where Ecologists Work: Biases in the Global Distribution of Terrestrial Ecological Observations." *Frontiers in Ecology and the Environment* 10: 195–201.

Mekonnen, A., C. Downs, E. O. Effiom, et al. 2021. "Can I Afford to Publish? A Dilemma for African Scholars." *Ecology Letters* 25: 711–715.

Miller, J., T. B. White, and A. P. Christie. 2023. "Parachute Conservation: Investigating Trends in International Research." *Conservation Letters* 16: e12947.

Mittermeier, R. A., R. A. Mittermeier, and S. A. C. V. Cemex. 2004. *Hotspots Revisited*. 1st English ed. Cemex.

Mori, A. S., S. Qian, and S. Tatsumi. 2015. "Academic Inequality Through the Lens of Community Ecology: A Meta-Analysis." *PeerJ* 3: e1457.

Mugal, M. A., S. A. Wani, F. A. Dar, et al. 2023. "Bridging Global Knowledge Gaps in Biodiversity Databases: A Comprehensive Data Synthesis on Tree Diversity of India." *Biodiversity and Conservation* 32: 3089–3107.

Myers, N., R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, and J. Kent. 2000. "Biodiversity Hotspots for Conservation Priorities." *Nature* 403: 853–858.

Nuñez, M. A., M. C. Chiuffo, A. Pauchard, and R. D. Zenni. 2021. "Making Ecology Really Global." *Trends in Ecology & Evolution* 36: 766–769.

Ocampo-Ariza, C., M. Toledo-Hernández, F. Librán-Embid, et al. 2023. "Global South Leadership Towards Inclusive Tropical Ecology and Conservation." *Perspectives in Ecology and Conservation* 21: 17–24.

Osuri, A. M., S. Machado, J. Ratnam, et al. 2020. "Tree Diversity and Carbon Storage Cobenefits in Tropical Human-Dominated Landscapes." *Conservation Letters* 13: e12699.

Poulter, B., L. Aragão, N. Andela, et al. 2019. "The Global Forest Age Dataset and Its Uncertainties (GFADv1.1)."

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Ramesh, B. R., M. H. Swaminath, S. V. Patil, et al. 2010. "Forest Stand Structure and Composition in 96 Sites Along Environmental Gradients in the Central Western Ghats of India: Ecological Archives E091-216." *Ecology* 91: 3118.

Riggio, J., J. E. M. Baillie, S. Brumby, et al. 2020. "Global Human Influence Maps Reveal Clear Opportunities in Conserving Earth's Remaining Intact Terrestrial Ecosystems." *Global Change Biology* 26: 4344–4356.

Rodgers, W. A., and H. S. Panwar. 1988. *Panning a Wildlife Protected Area Network in India*. Vol. 1. Wildlife Institute of India.

Srinivasan, U., N. Velho, J. S. H. Lee, D. D. Chiarelli, K. F. Davis, and D. S. Wilcove. 2021. "Oil Palm Cultivation Can Be Expanded While Sparing Biodiversity in India." *Nature Food* 2: 442–447.

Srivathsa, A., D. Vasudev, T. Nair, et al. 2023. "Prioritizing India's Landscapes for Biodiversity, Ecosystem Services and Human Well-Being." *Nature Sustainability* 6: 568–577.

Stocks, G., L. Seales, F. Paniagua, E. Maehr, and E. M. Bruna. 2008. "The Geographical and Institutional Distribution of Ecological Research in the Tropics." *Biotropica* 40: 397–404.

Sullivan, M. J. P., J. Talbot, S. L. Lewis, et al. 2017. "Diversity and Carbon Storage Across the Tropical Forest Biome." *Scientific Reports* 7: 39102.

Tewari, V. P. 2016. "Forest Inventory, Assessment, and Monitoring, and Long-Term Forest Observational Studies, With Special Reference to India." *Forest Science and Technology* 12: 24–32.

Tinkham, W. T., P. R. Mahoney, A. T. Hudak, et al. 2018. "Applications of the United States Forest Inventory and Analysis Dataset: A Review and Future Directions." *Canadian Journal of Forest Research* 48: 1251–1268.

Tiruvaimozhi, Y. V., J. Borah, C. P. Kala, et al. 2025. "Long-Term Ecological Monitoring in India: A Synthesis." *Biological Conservation* 310: 111344.

Venter, O., E. W. Sanderson, A. Magrach, et al. 2016. "Global Terrestrial Human Footprint Maps for 1993 and 2009." *Scientific Data* 3: 160067.

Wickham, H., M. Averick, J. Bryan, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4: 1686.

Wickham, H., R. François, L. Henry, and K. Müller. 2022. "Dplyr: A grammar of Data Manipulation."

Wilkinson, M. D., M. Dumontier, I. J. J. Aalbersberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 160018.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Appendix S1:** Detailed methods for screening INvenTree metadataset. **Appendix S2:** INvenTree data spread on other base layers. **Appendix S3:** Records of studies where corrections were applied. **Appendix S4:** Additional methods for analyzing FAIR principles in INvenTree metadataset. **Appendix S5:** Supplementary results for authorship analysis. **Appendix S6:** Additional methods for sampling priority maps. **Appendix S7:** Sampling priority maps using alternate categorisations.