Pervasive gaps in Amazonian ecological research

Graphical abstract

Highlights

- Ecological metadata were compiled for 7,694 sites across the Brazilian Amazon
- Accessibility and proximity to research facilities influenced research probability
- Knowledge gaps are greater in uplands than in wetlands and aquatic habitats
- Undersampled areas overlap predicted hotspots of climate change and deforestation

Authors

Raquel L. Carvalho, Angelica F. Resende, Jos Barlow, ..., Jansen Zuanon, The Synergize Consortium, Joice Ferreira

Correspondence

raqueluly@gmail.com (R.L.C.), gel.florestal@gmail.com (A.F.R.), jos.barlow@lancaster.ac.uk (J.B.), mariormoura@gmail.com (M.R.M.)

In brief

Carvalho et al. map the locations and drivers of ecological research across the Brazilian Amazon. Research facilities and accessibility were strong predictors of research location. Areas with the lowest probability (<0.1%) of research covered about 27.3%, 17.3%, and 54.1% of aquatic, wetland, and upland habitats, respectively.
Pervasive gaps in Amazonian ecological research


1Empresa Brasileira de Pesquisa Agropecuária, Amazônia Oriental, Belém 66095-903, Brazil
2Universidade de São Paulo, São Paulo 05508-220, Brazil
3Universidade de São Paulo, Esalq, Piracicaba 13418-900, Brazil
4Lancaster University, LA1 4YQ Lancaster, UK
5University of Bristol, BS8 1QU Bristol, UK
6Universidade Estadual de Campinas, Campinas 13083-862, Brazil
7Universidade Federal da Paraíba, Areia 58397-000, Brazil
8Universidade de Brasília, Brasília 70910-900, Brazil
9Universidade do Porto, 4099-002 Porto, Portugal
10University of Oxford, OX1 3QY Oxford, UK
11University Federal Rural da Amazônia, Belém 66077-830, Brazil
12Universidade Federal de Mato Grosso, Cuiabá 78060-900, Brazil
13Universidade Federal do Rio Grande do Norte, Natal 59078-970, Brazil
14Universidade Federal do Amazonas, Manaus 69067-375, Brazil
15Instituto Nacional de Pesquisas da Amazônia, Manaus 69067-005, Brazil
16University of Oxford, OX1 3QY Oxford, UK
17Virginia Tech, Blacksburg, VA 24061, USA
18Universidade Federal do Amazonas, Belém 66077-830, Brazil
19University of Leeds, LS2 9JT Leeds, UK
20Universidade Federal do Acre, Rio Branco 69920-900, Brazil
21Naturalis Biodiversity Center, 2333 CR Leiden, the Netherlands
22Utrecht University, 3584 CS Utrecht, the Netherlands
23Universidade Federal de Mato Grosso, Cuiabá 78060-900, Brazil
24Universidade Federal do Rio Grande do Norte, Natal 59078-970, Brazil
25Museu Paraense Emílio Goeldi, Belém 66040-170, Brazil
26These authors contributed equally
27Lead contact

*Correspondence: raqueluly@gmail.com (R.L.C.), gel.forestal@gmail.com (A.F.R.), jos.barlow@lancaster.ac.uk (J.B.), mariormoura@gmail.com (M.R.M.)

https://doi.org/10.1016/j.cub.2023.06.077

SUMMARY

Biodiversity loss is one of the main challenges of our time,1,2 and attempts to address it require a clear understanding of how ecological communities respond to environmental change across time and space.3,4 While the increasing availability of global databases on ecological communities has advanced our knowledge of biodiversity sensitivity to environmental changes,5–7 vast areas of the tropics remain understudied.8–11 In the American tropics, Amazonia stands out as the world’s most diverse rainforest and the primary source of Neotropical biodiversity,12 but it remains among the least known forests in America and is often underrepresented in biodiversity databases.13–15 To worsen this situation, human-induced modifications16,17 may eliminate pieces of the Amazon’s biodiversity puzzle before we can use them to understand how ecological communities are responding. To increase generalization and applicability of biodiversity knowledge,18,19 it is thus crucial to reduce biases in ecological research, particularly in regions projected to face the most pronounced environmental changes. We integrate ecological community metadata of 7,694 sampling sites for multiple organism groups in a machine learning model framework to map the research probability across the Brazilian Amazonia, while identifying the region’s vulnerability to environmental change. 15%–18% of the most neglected areas in ecological research are expected to experience severe climate or land use changes by 2050. This means that unless we take immediate action, we will not be able to establish their current status, much less monitor how it is changing and what is being lost.
RESULTS

Our detailed assessment of the ecological research in the Brazilian Amazon assessed how logistics and human influence on the forests explained research probability across 7,694 community ecology sites surveyed from 2010 to 2020. Across nine organism groups—benthic invertebrates, heteropterans, odonates, fishes, macrophytes, birds, woody vegetation, ants, and dung beetles—ecological research was unevenly distributed in all three major habitat types investigated with easily identifiable research gaps (research probability < 0.1) covering 54.1% of unflooded areas locally known as terra firme (uplands, hereafter), 27.3% of aquatic habitats, and 17.3% of wetlands (Figure 1). While ecological research effort differs across organism groups, our findings highlight very consistent spatial patterns of research probability across habitat types, even among groups assessed in multiple habitat types (plants and birds; Figure 2).

Drivers of research biases

Overall, logistics and human influence factors explained 64% of the variation in research probability. Among the logistic-related factors, accessibility and distance to research facilities consistently emerged as important predictors of research probability (Figure 3), highlighting the role of logistical constraints and ease of access. Research probability increased with closer proximity to transportation and research facilities for all upland organisms and most representatives of wetlands and aquatic habitats. In addition, dry season length mattered for ecological research on wetland birds but showed little contribution with respect to other organisms (Figure 3C). Dry season length also had a contrasting effect across habitat types, increasing the research probability in uplands and aquatic habitats, but decreasing it in wetlands. Although logistics influenced ecological research the most, forest degradation and land tenure also showed a modest but consistent importance across all organism groups. Both predictors affected ecological research in the same direction across organisms, with research probability slightly declining in more degraded areas and indigenous lands but increasing in protected areas (Figures 3D and 3E).

Research biases and projected environmental changes

Unfortunately, about half of the Brazilian Amazon is either already deforested (23.50%) or projected (27.29%) to be by 2050, with these regions showing contrasting chances of
ecological research (Figures 4E and S2). For instance, research probability was higher among areas currently deforested than in areas projected to be deforested within the next three decades (Figure 4E), corroborating the finding that research has occurred mostly in human-modified landscapes. Our findings also indicate that 15%–18% of the most scientifically neglected areas (Figure S2), herein defined as those in the first quartile of research probability, show high susceptibility to climate changes by 2050 (Figures 4A and 4C) and habitat destruction (Figures 4B and 4D).

**DISCUSSION**

We elucidate how logistics and human influence have affected ecological research coverage in Brazilian Amazonia. Our comprehensive regional synthesis improves the understanding of Amazonian ecological research and opens new avenues to redress the underrepresentation of tropical rainforests in biodiversity databases. Besides offering further support for the role of accessibility and research facilities in ecological studies, our findings also highlight some of the challenges involved in expanding research to areas that have not been sampled before. While increasing overall research effort is valuable for many reasons, our models show that this would mostly lead to more surveys in areas with a high probability of research, as the lack of accessibility and research infrastructure promotes close distances between new and existing sampling sites. Additional measures will be required to overcome these barriers and reach regions where the research probability is low. Here we explore two key challenges related to resolving spatial research gaps in the Brazilian Amazon.

**Remote regions**

Accessibility was a key factor in our results. One strategy to reach undersampled regions would be to fund expedition-like programs such as the **Expedição Serra da Mocidade**, where Amazon-based researchers from different institutions surveyed remote mountains in northern Amazonia, where no previous studies had been conducted. While the **Expedição Serra da Mocidade** focussed mainly on finding new species and understanding species distributions, an ecological expedition would require longer periods in the field to enable the use of the standardized sampling protocols required to assess biotic changes across space and time. For instance, regions with low research probability partially overlap with those projected to experience either high and low climate change, as well as with regions facing high risk of future deforestation and degradation (Figure 4). Hence, enhancing ecological research in these remote regions could be the sole chance to unveil pieces of the Amazon’s biodiversity puzzle before they succumb to human-induced modifications, while also seizing one of our prime opportunities to comprehend climate change effects without the potential anthropogenic influence.

Distance to research centers was also an important factor, and an alternative approach would be to fund new centers in cities that are within or close to the areas with the lowest probability of research. While this could be a more logistically demanding approach, it has four major longer-term benefits. First, it would encourage the training of local researchers and help science to endure beyond individual assessments. Second, it could build the base for more detailed scientific research, including the logistical support necessary to identify and prepare specimens and develop local collections. Third, it could enhance the capacity required to conduct the longer-term research required to understand global change. Finally, it may be a more sustainable strategy than investing in expeditions, as these are rarely repeated and risk being canceled under new governments. Although research centers have suffered under recent governments, being subjected to long-lasting neglect affecting funding, structural maintenance, and emptying of key scientific and technical staff, they have also shown themselves to be resilient, and continue to lead and support long-term research across the Amazon. Whatever approach is taken, it is key that it does not undermine the current network of research and education facilities in the Brazilian Amazon or remove funding from long-term monitoring programs on biodiversity baselines and ecological changes, which often take place near existing research facilities.

**Indigenous lands**

Results highlighted the limited research effort in indigenous lands when compared to strictly protected and sustainable use reserves. This reflects a major ecological knowledge gap, as indigenous lands cover about 23% of the Brazilian Amazon and play a fundamental role in preserving Amazonia’s biocultural diversity. Over the last years, Brazil’s indigenous lands have come under increasing threat from illegal activities, such as logging, invasion by squatters, and gold mining, with this latter activity also strongly impacting the health of riverside peoples. The lack of government support has forced traditional and
indigenous peoples to defend their territories on their own.\(^\text{30}\)

Indigenous lands such as the Kaiapó and Arariboia already represent some of the last areas of extensive forest in the south and eastern Amazon, and the role of indigenous lands is expected to become more pronounced under business-as-usual deforestation practices (Figure 4E).
While bureaucratic requirements and limitations in local communicability may reduce the research propensity in such areas, coordinated actions between the environment ministry (a specific organization within the Brazilian government) research centers and indigenous peoples have high potential to minimize knowledge gaps within indigenous lands. Any such knowledge co-production needs to be equitable and decolonial, and to recognize and respect the diverse knowledge systems in place. For instance, traditional practices of indigenous people should be valued and incorporated into research methodologies, ensuring their active participation and ownership in the process. The co-creation of research with local communities will likely require additional support and training for both the scientific community and for local communities and their organizations. Such a program could have many additional benefits, as enhancing the involvement of local communities could support more inclusive science as well as better resource management and livelihoods.

Differences between organisms and habitats
We find remarkably consistent drivers of research effort across different organisms and habitats. But one noticeable exception is the disproportionate importance that a small number of local experts have in the spatial sampling of underrepresented taxa. For instance, most sampling (95%) of aquatic invertebrates (hemipterans and odonates) are distributed in eastern Amazonia (Pará state) and come from a single research group composed of local experts based in the Amazonian city of Belém. Other aquatic invertebrate specialists have collected in different regions of Amazonia, but their focus on taxonomy means they rarely use the standardized sampling required for ecological data. Considering the distinct objectives of taxonomic research and ecological sampling, prior planning and greater collaboration will be required to achieve the potential benefits that could be accrued from integrating Amazonian ecology and taxonomy. We also demonstrate that research gaps are higher for uplands than wetlands and aquatic habitats, which likely reflects the role that the broad network of navigable waterways has in facilitating access to wetland and aquatic areas.

Can regional curation help reduce global biases?
The metadata we based this research on describe a large number of datasets with substantial coverage across multiple organisms in the Brazilian Amazon (see the Synergize project). To date, few of these datasets are integrated into global databases. For example, the BioTIME, BIOFRAG, FragSAD, and Predicts databases collectively include only 222 datasets for the...
Conclusion
Our large-scale assessment of ecological research across the Brazilian Amazon not only highlights the extent of research gaps and biases in tropical systems, but also provides valuable insights into potential solutions to improve conservation planning for the world’s most diverse rainforest. We show the importance of going beyond areas that are accessible and close to research bases, and expanding research into regions that will likely be affected by climate change or deforestation. Doing so will not be easy, and ecology alone will not help resolve the environmental crises facing the world. But understanding the responses of biodiversity and ecosystems forms a key part of keeping society informed about its impacts and supporting the implementation of evidence-based policies and practices that can help mitigate the worst outcomes.

STAR METHODS
Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Data collection
- METHOD DETAILS
  - Predictors of research probability
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Research probability modelling
  - Overlap between research and anthropogenic disturbances
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION
Supplemental information can be found online at https://doi.org/10.1016/j.cub.2023.06.077.

CONSORTIA

AUTHOR CONTRIBUTIONS


DECLARATION OF INTERESTS

We support inclusive, diverse, and equitable conduct of research.

INCLUSION AND DIVERSITY

The authors declare no competing interests.

REFERENCES


STAR METHODS

KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deposited data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw data and code for the analyses on research probability across the Brazilian Amazoni</td>
<td>This paper</td>
<td><a href="https://zenodo.org/record/7951033">https://zenodo.org/record/7951033</a></td>
</tr>
<tr>
<td>Software and algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R software</td>
<td>The R Foundation / R Development Core Team</td>
<td><a href="https://www.r-project.org">https://www.r-project.org</a></td>
</tr>
<tr>
<td>Rstudio IDE</td>
<td>AGPL v3 / Posit-PBC</td>
<td><a href="https://posit.co/products/open-source/rstudio/">https://posit.co/products/open-source/rstudio/</a></td>
</tr>
<tr>
<td>Google Earth Engine</td>
<td>Google LLC</td>
<td><a href="https://earthengine.google.com/">https://earthengine.google.com/</a></td>
</tr>
<tr>
<td>QGIS</td>
<td>QGIS Development Team</td>
<td><a href="https://www.qgis.org/en/site/">https://www.qgis.org/en/site/</a></td>
</tr>
<tr>
<td>Inkscape</td>
<td>GNU GPL-2.0-or-later</td>
<td><a href="https://inkscape.org/">https://inkscape.org/</a></td>
</tr>
</tbody>
</table>

RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Raquel L. Carvalho (raqueluly@gmail.com).

Materials availability
This study did not generate new reagents, sequences or eventuate in the archiving of specimens.

Data and code availability
- Metadata have been deposited at Zenodo and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Data collection
All data collection took place within the SYNERGIZE (Synthesizing Ecological Responses to Degradation in Amazonian Environments) project, a collaborative effort of scientists from institutions in Brazil and other countries that integrate and synthesise data on Amazonian biodiversity studies. More information about the project and study region can be found elsewhere (key resources table).

We gathered ecological community datasets for nine organism groups (benthic invertebrates, heteropterans, odonates, fishes, macrophytes, birds, woody vegetation, ants, and dung beetles) between 2010 and 2020 in the Brazilian Amazonia. We focused on this period to keep results relevant to the recent potential drivers of research effort.

To be included in our database, any published research needed to be considered as an ecological community dataset derived from quantitative and repeatable sampling protocols, with sampled taxa identified at a minimum of family level. For each selected study (see Table S1), we contacted the first and/or corresponding author. Because work published in English may be a biased research subset, we extended our search by contacting experts known to work in the Brazilian Amazonia. In the following, we have summarized the information derived from the compiled data (but see Table S1).

Woody vegetation
We first compiled the metadata available in three consolidated databases: ForestPlots.net, Amazon Tree Diversity Network (Rainfor), and Secondary Forests Research Network (2ndFOR) for Brazil. These databases have deployed huge efforts to acquire data and metadata across the Brazilian Amazonia. For the ForestPlots database, we contacted 79 people to request data owners’ permission to use the metadata. In 1% of the number of people contacted, we had no answers from any team member and, consequently, no permission to use metadata. In parallel, we listed all the universities and research institutes with graduate courses in ecology or related fields in all states covering the Brazilian Amazonia, and invited additional
90 data owners to collaborate with their metadata (Table S1). Overall, 47 authors returned our contact and shared metadata on 2,597 inventories meeting our criteria.

**Terrestrial animals**
We focused on published studies on ants, birds, dung beetles in Amazonia using the Web of Science platform (Table S1 for string words used and search dates). Searches were conducted in English and with no restriction on year. We did not focus on a search period between 2010 and 2020 for these groups, as the Synergize project is developing a database that goes beyond the period considered here TAOCA. From the initial total of 2,244 published manuscripts obtained through the platform, only 225 were ecological studies. Among these 225 studies, we contacted 99 authors, many of whom were corresponding authors of more than one article. Ninety one percent of these researchers returned our contact and shared metadata of their studies (Table S1).

**Aquatic groups**
We focused on published studies on Amazon fishes using the Web of Science platform (Table S4 for string words used and search dates). Since fishes are also research subjects in many areas of knowledge (such as social food sciences), we selected only articles related to biodiversity and conservation areas. For other aquatic groups (benthic, heteropterans, odonates and macrophytes), we focused on a period between 2010 and 2020. We divided aquatic invertebrates into benthic, heteropterans, odonates, because most datasets on heteropterans and odonates referred to sampling of adults near streams, while for benthic invertebrates (Ephemeroptera, Plecoptera, Trichoptera, Diptera and Coleoptera), the datasets corresponded to the sampling of larvae in streams. From the initial total of 1,073 published manuscripts obtained through the platform, only 523 were ecological studies. Among these 523 studies, we contacted 69 authors, many of whom were corresponding authors of more than one article. Eighty-one percent of these researchers returned our contact and shared metadata of their studies (Table S1).

All sampling sites – informed within the ecological studies herein considered – had their geographic coordinates verified and cross-checked with databases of the political and environmental limits of the Brazilian Amazon. Studies were included if data collection utilized repeatable methods, which varied among different biological groups (see details in Table S1). Organism groups were sampled along transects, grids, or plots, which exhibited significant variations in size.

Our original dataset comprised 3,917 sampling sites in uplands (635 for ants, 742 for birds, 687 for dung beetles, and 1,853 for woody vegetation), 923 in wetlands (18 for ants, 136 for birds, 25 for dung beetles, and 44 for woody vegetation) and 2,897 in aquatic habitats (536 for benthic invertebrates, 1710 for fishes, 153 for heteropterans, 159 for macrophytes and 339 for odonates), totalling 7,737 sampling sites.

Metadata used in this study, including habitat, year of study, organism groups, and geographical coordinates (longitude and latitude) can be found in the STAR Methods (see data and code availability). Other metadata and community data can be found in the original sources, which varied between groups. For woody vegetation, data is available in three consolidated databases: ForestPlots, Amazon Tree Diversity Network (Rainfor, ATDN), and Secondary Forests Research Network (2ndFOR). For terrestrial fauna, the data are deposited at the TAOCA platform. Fish group data is available in the AmazonFish, aquatic invertebrates and macrophytes are available, under request, with Leandro Juen and Thaise Michelan, respectively, both from the Laboratory of Ecology and Conservation (UFPA, Brazil).

**METHOD DETAILS**

**Predictors of research probability**
To model research probability for each organism group, we used predictors related to logistics (accessibility, research facilities, and dry season length), and human influence (land tenure, degradation, and dry season length). We define human influence as those variables used to assess the ecological condition of a particular habitat or ecosystem. We used logistics-related variables as certain areas are easier to work in due to accessibility, proximity to research centres, or location within public protected areas. As the dry season length affords extended accessibility to many of the drier regions in the Brazilian Amazon each year, we deemed it a crucial logistical factor. Although predictor variables change over time, we assumed that these changes in a short time window are negligible relative to the variation across space and adopted a static layer to represent the Brazilian Amazonia over the ten years from which data were collected.

**Accessibility**
We used travel time from each surveyed site to the nearest village or city, considering a population between five thousand to five million people in 2015. The travel time to the nearest major city indicates the fastest travel speed considering all transportation facilities (road, river, canal, and sea lane) as well as characteristics of land cover, slope, and elevation of the areas. We used the dataset available in Weiss and collaborators, downloaded at the spatial resolution of 30 arc-sec (Figure S1A).

**Research facilities**
We used the geographical location of institutes and universities that offer undergraduate and graduate courses related to biology, ecology, and forestry, as well as experimental farms, forest reserves, and other institutions that support these courses. We obtained the geographical distribution of undergraduate courses from the Brazilian Education Ministry, and we extracted information on graduate courses from the Coordination for the Improvement of Higher Education Personnel (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES). We also gathered data on institutes from the Ministry of Agriculture, Livestock and Supply (Ministério da Agricultura, Pecuária e Abastecimento – MAPA) and Ministry of the Environment and Climate Change (Ministério do Meio Ambiente – MMA). In addition, we gathered information on experimental farms and forest reserves from the website of each
University or Research Institution, covering all Brazilian states that spatially overlap with Amazonia. Then, we rasterised this predictor at 8 arc-sec (~0.25 km) resolution. Finally, we extracted the distance from each surveyed site to the nearest university or infrastructure containing research facilities (research facility hereafter) (Figure S1B).

**Dry season length**

We used the average dry season length in consecutive months, extracted from the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) dataset. The CHIRPS dataset was created from interpolation techniques of satellite data and rainfall station data from 1981 to the present. We considered dry months to be those with less than 100 mm of precipitation. This variable ranged from zero to eight months and originally presented a 2.5 arc-min (~5 km) resolution (Figure S1D).

**Degradation**

We extracted the accumulated degradation from 1995 to 2017 using a 1 arc-sec (~0.03-km) resolution layer developed by Bullock and collaborators. This is the most suitable and recent dataset for the period considered in this study (but see Matricardi et al.). However, it does not cover a small area of our study – the Amazon Network of Georeferenced Socio-Environmental Information-RAISG – which led us to redefine our study area limits according to Bullock and collaborators. Therefore, 3.2% of sampling sites in our original database were removed from analyses described below as they fell outside the RAISG-defined Amazonia biome (see Figure S1F). Degradation was considered as any natural or anthropogenic disturbance (e.g., fire, windthrow, selective logging, and fuelwood) that does not alter a pixel’s original land cover forest category. To assess the forest degradation in the vicinity of the sampling area, we counted the proportion of degraded forest in a circle with a radius of 1000 m surrounding each plot. We used 1 km as a compromise between capturing enough forests to get a reasonable degradation sample while minimising the overlap between the buffers of neighbouring plots. The mean geographic distance between plots was 926.9 km (range: 0–2536.3 km, Figure S1E).

**Land tenure**

We obtained a land tenure map at 1 arc-sec (~0.03-km) of spatial resolution for the year 2017. We classified the surveyed sites in nine categories: (1) Indigenous lands, both unofficial and official land by the National Foundation for Indigenous People; (2) Unassigned public land that has not been registered in National Institute of Colonisation and Agrarian Reform (INCRA) database and/or incorporated into other public land categories; (3) Private land recognised by public government databases [e.g. INCRA and Environmental Rural Registry]; (4) Protected area, with sustainable use of natural resources [i.e. low-level non-commercial use of resources allowed [category VI, IUCN]]; (5) Strict Reserve area, where the use and management of natural resources are strictly controlled and limited to science or wilderness preservation [category IA, IUCN]; (6) Rural settlements registered by INCRA; (7) Water body, including continental and/or coastal waters; (8) Quilombola land, afro-Brazilian settlements established by escaped or enslaved people up to the 19th century; and (9) Other, including railway, road, military and urban areas (Figure S1C).

**QUANTIFICATION AND STATISTICAL ANALYSIS**

**Research probability modelling**

Before running the analyses, we defined the habitats (upland, wetland and aquatic) using existing forest/non-forest and wetland layers. We removed the non-forested areas from our study areas using a 25 m resolution ALOS PALSAR 2 Forest/Non-Forest map for 2015. Next, we obtained the wetland delimitation by complementing the Hess and collaborators layer, which covers the lowland Amazon Basin wetlands (below 500 m asl), with the wetlands layer from the Center for International Forestry Research (CIFOR) to cover higher altitudes and eastern Amazon Forest areas outside the basin (Figure S1F). For the wetland habitat, we removed upland and water surface, keeping only forested areas. Lastly, we used the same wetland delimitation for aquatic habitat but included open water bodies by joining a river level layer. We removed only the areas that did not meet the inclusion criteria (being the habitat where each taxon occurs and being forest for upland and wetland habitats; Figure S1F). We considered a 1 km spatial resolution and the extension of the degradation layer to run all analyses. In addition, we used equal-area projections for all our layers (CRS WGS 84 – EPSG: 4326).

We used occurrence data of community ecological studies (i.e., surveyed sites) in a Random Forest (RF) framework to model the ecological probability of sampling research for each biological group in each pixel (1 km) across the Brazilian Amazon. Before building RF models, we tested for multicollinearity among predictors (accessibility, research facility, land tenure, dry season length, and degradation layers) using the Variance Inflation Factor (VIF). Computations were performed separately for each combination of biological group and habitat type. Since all predictors showed VIF < 2, we retained each variable in the analyses (Table S2). We undertook VIF analyses using ‘usdm’ package.

We built RF models of ecological research probability (hereafter, research probability) separately for each organism and habitat. Like many species distribution modelling techniques, RF requires absence or pseudoabsence data, and shows better performance when used with balanced classes (i.e., pseudoabsence ratio of 1:1). Therefore, we computed pseudoabsences using the same number of observed presences in each organism group. We validated the RF models using a 5-fold cross-validation approach. The presence-pseudoabsence data were randomly partitioned into five equal parts, with four of those used for training and the fifth as validation-fold. For each RF model, we initially used the `tuneRF` function in the randomForest package to identify the number of predictors (ntree argument) and trees (ntree argument) returning the most accurate RF, that is, the model with the lowest mean-squared error (MSE) and highest explained deviance (R²). RF computations were then repeated for each training-fold using the optimal parameters identified by `tuneRF`. We excluded ants and dung beetles from the RF models for wetland analyses, since they had fewer than 25 sampling sites.
To evaluate model performance, we used the Sørensen similarity index to measure the similarity between predictions and observations. This index is independent of the prevalence of sampled sites (ratio of observed presences to all sites) and less sensitive to under/overprediction issues than metrics based on specificity, such as the True Skill Statistic – TSS. We measured variable importance using the proportional increase in MSE, which measures the relative decrease in model accuracy by shuffling variable values. We also built partial dependence plots to represent the marginal effects of predictor values on research probability across taxa and habitats.

Overlap between research and anthropogenic disturbances
We intersected research probability with susceptibility to current and future anthropogenic disturbances to identify areas with ecological knowledge most at risk. We used three indicators of susceptibility to anthropogenic disturbances that reflect continuous trends of long-term demographic growth and economic development: climate change, deforestation, and degradation.

To indicate climate change, we computed the difference between current and future projections of climate represented through 13 bioclimatic variables (e.g., $\Delta$Bio1 = Bio1$_{\text{future}}$ − Bio1$_{\text{current}}$) obtained from the Intergovernmental Panel on Climate Change (IPCC) Interactive Atlas (Table S4). For each bioclimatic variable, the $\Delta$climatic values were measured between projections for 2041-2060 and 1981-2010 under the SSP585 scenario. To provide a consensus metric of future climate change across different generalised circulation models (GCM), we average $\Delta$climatic values across all GCM available at the IPCC Interactive Atlas (Table S4). We rescaled the $\Delta$climatic values in the interval between -1 and 1 and passed their absolute values ($|\Delta\text{climate}|$) through a Principal Component Analysis (PCA) to remove multicollinearity. The magnitude of climate change for each pixel was calculated as the Euclidean distance between its position in the two first axes PCA space and the origin; which represents a reference point of no climate change ($|\Delta\text{climate}| = 0$ along all axes).

To identify areas most threatened by deforestation and degradation, we used projections made recently available indicating trends under a business-as-usual scenario for 2050. The term deforestation refers to the complete removal of canopy cover, whereas degradation is the term used to describe a natural or human-induced disturbance that does not alter the land cover category assigned to a pixel. Currently deforested areas are also shown since they may be reforested in case of land abandonment or political incentives. We prepared the spatial layers using Google Earth Engine (GEE) and carried out all statistical analyses in R 4.0.5.

Next, both measures of anthropogenic disturbances, (i) magnitude of climate change and (ii) combined deforestation and degradation, as well as research probability, were split into equal-sized quantiles holding 0–25%, 25–50%, 50–75%, and 75–100% of samples (pixels). We used the first quartile of research probability (0–25%) to identify the most neglected areas in ecological research, and the last quartile (75–100%) of climate change and deforestation-degradation to identify areas most susceptible to anthropogenic disturbances.

ADDITIONAL RESOURCES
For terrestrial fauna, the metadata used in this study were stored in the TAOCA database. This database emerged in response to the demand for organizing, standardizing, and securely storing a large amount of data received by the Synergize project (https://www.taoca.net/).