# Data availability impacts the predictive accuracy of pressure-based biodiversity models

Jakob Nyström[1,2*], Jeffrey R. Smith[3,4], Lisa Mandle[5], Andrew Gonzalez[6–8], Thomas B. Schön[9], Tobias Andermann[1,2]

1. Biodiversity Data Lab, Department of Organismal Biology, Uppsala University, Uppsala, Sweden.
2. Science for Life Laboratory, Uppsala University, Uppsala, Sweden.
3. Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, USA.
4. High Meadows Environmental Institute, Princeton University, Princeton, New Jersey, USA.
5. Natural Capital Project, Stanford University, Stanford, California, USA.
6. Department of Biology, McGill University, Montreal, Quebec, Canada.
7. Québec Centre for Biodiversity Science, Montreal, Quebec, Canada.
8. Group on Earth Observations Biodiversity Observation Network, Montreal, Quebec, Canada.
9. Department of Information Technology, Uppsala University, Uppsala, Sweden.

[*]Corresponding author: Jakob Nyström, jakob.nystrom@ebc.uu.se

## Abstract

Amidst the biodiversity crisis, there is high demand for spatially explicit biodiversity indicators. Global models that quantify impacts of human pressures provide important insights for conservation, but their accuracy in spatial projections has yet to be systematically tested. Here we evaluate this using a global dataset of 25,987 species inventories from 681 studies. We find that mixed models with study attributes as random effects – common in meta-analysis and used in several indicators – exhibit low predictive accuracy. This is driven by reliance on a small set of averaged fixed effects. In contrast, a biogeographic-taxonomic model structure with explicit environmental covariates shows relatively higher interpolation accuracy. However, accuracy when extrapolating to other contexts remains low, due to distribution shifts in environmental conditions. These patterns apply to site-level diversity and differences between sites. Both models estimate similar land-use impacts, in line with previous research, yet our results highlight the challenging gap between effect size inference and prediction. Models are essential for informed conservation efforts, but their applicability is fundamentally constrained by data availability. Whereas countries with extensive data can build high-fidelity national indicators, accelerated data collection and model development are needed to better support data-poor regions with localized and actionable insights.

# Introduction

Terrestrial biodiversity is declining on a global scale[1–3], caused by land use change, natural resource exploitation, pollution, climate change, and invasive species[3,4]. Biodiversity loss threatens the stability of ecosystems and the services they provide, on which human prosperity depends[5]. The recently updated monitoring framework[6] of the Global Biodiversity Framework[7] (the GBF-MF) underscores the critical role of indicators to halt and reverse this development[8]. In parallel, companies are accelerating efforts to understand nature-related dependencies, impacts, and risks[9–11]. Demand is therefore growing for robust, globally consistent indicators for assessing biodiversity change and its drivers[12].

While global biodiversity repositories contain increasing amounts of data, lingering geographic and taxonomic gaps make comprehensive biodiversity monitoring challenging[13–15]. Statistical models that quantify how biodiversity correlates with environmental drivers and human pressures[16] have the potential to fill data gaps through projections. Models could thereby support national monitoring[6], global assessments[2], and scenario analysis[1,17,18], presenting a forward-looking complement to the GBF-MF headline indicators[16,18]. However, modeling biodiversity dynamics at large scales raises questions about the generality of models: the extent to which inferences from a sample apply to the sampled ecological context at large, as well as to other contexts[19]. Generality is a major challenge when estimating effect sizes of drivers[19], let alone when extending parameter inference to spatially explicit predictions of biodiversity. Ecological prediction is notably hard due to local dynamics emerging from a combination of evolutionary patterns, biotic and abiotic interactions, scale dependencies, and biological stochasticity[20].

Challenges notwithstanding, scalable monitoring of biodiversity change and its drivers are key questions for biodiversity research[16]. In this context, biodiversity is measured relative to some baseline condition in time or space, often using the concept of biodiversity intactness: 'the average abundance of a large and diverse set of organisms in a given geographical area, relative to their reference populations'[21]. The intactness of a given location can be quantified using pressure-response models of species richness, abundance and composition, in relation to reference sites that are relatively ecologically intact. Global model-based intactness indicators, such as the Biodiversity Intactness Index (BII)[22,23] and Mean Species Abundance (from GLOBIO)[24], have been widely adopted, for example by the GBF-MF[6], the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES)[2], and the private sector[12]. This adoption underscores the value of biodiversity models to support policy and decision-making[8,12,16,18].

Efforts to develop large-scale, model-based indicators have advanced biodiversity modeling, generating important insights about the effects of human pressures[22,24] with major value for conservation policy. Yet, there has also been critique against lack of agreement with other global metrics[25], limited applicability in data-poor regions[26], and insufficient model testing[25]. Considering their applied relevance, the use of pressure-response models for spatially explicit projections calls for more systematic evaluation of predictive performance. Evaluation is particularly important in this context, since data for global models relies on aggregation of many heterogeneous source studies into meta-databases. The spatial and taxonomic heterogeneity in study contexts, combined with gaps in underlying data, imply a need for extensive extrapolation to produce continuous output maps. Results from macroecological and species distribution modeling show that reasonable model performance can often be achieved when predicting in a context similar to the training data, while accuracy is typically much lower when predicting into new contexts[27–29]. However, such results are generally based on much narrower contexts than the

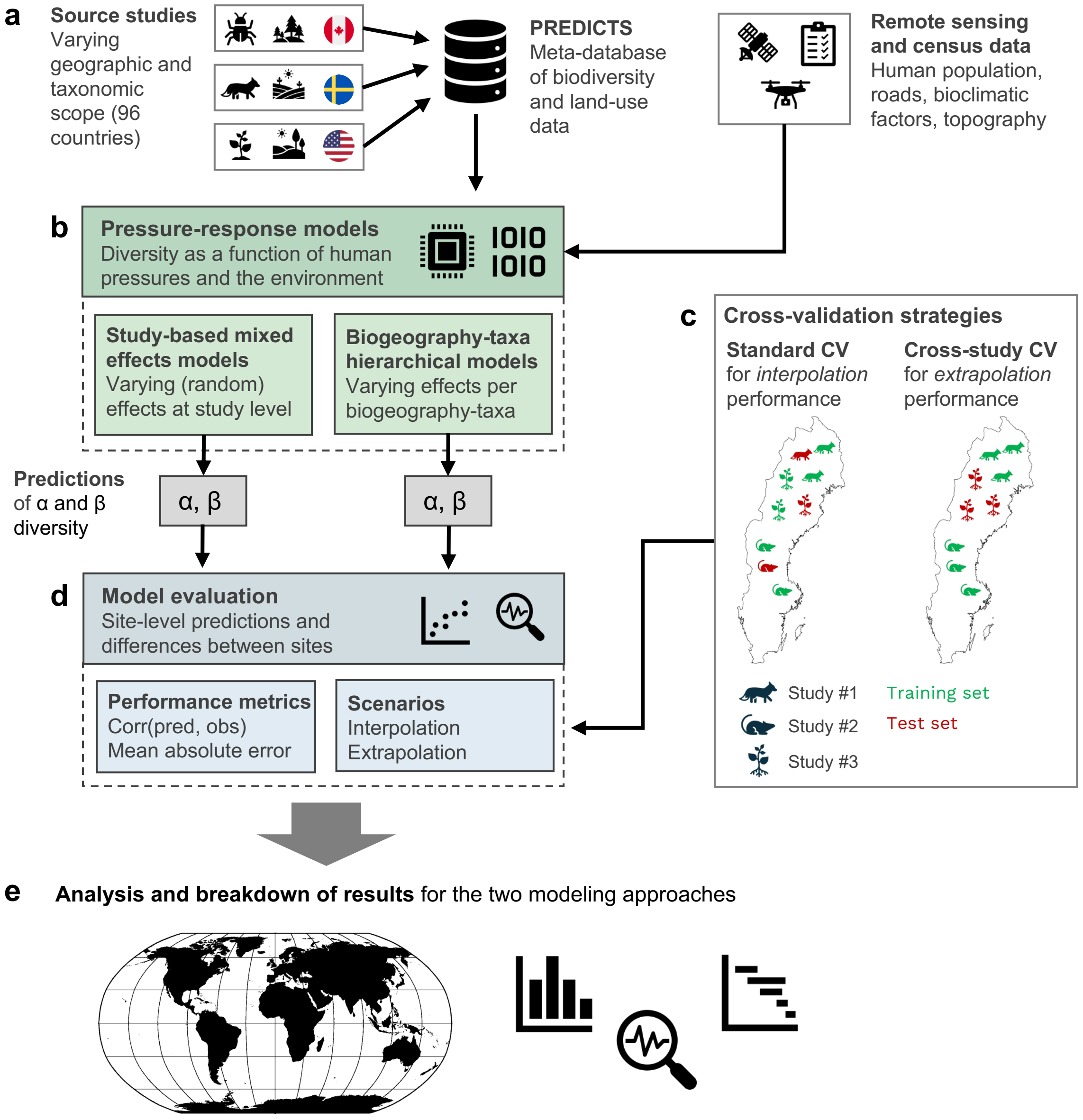

a
Source studies
Varying geographic and taxonomic scope (96 countries)
PREDICTS
Meta-database of biodiversity and land-use data
Remote sensing and census data
Human population, roads, bioclimatic factors, topography
b
Pressure-response models
Diversity as a function of human pressures and the environment
1010
1010
Study-based mixed effects models
Varying (random) effects at study level
Biogeography-taxa hierarchical models
Varying effects per biogeography-taxa
Predictions of α and β diversity
α, β
α, β
c
Cross-validation strategies
Standard CV
for *interpolation* performance
Cross-study CV
for *extrapolation* performance
Study #1
Study #2
Study #3
Training set
Test set
d
Model evaluation
Site-level predictions and differences between sites
Performance metrics
Corr(pred, obs)
Mean absolute error
Scenarios
Interpolation
Extrapolation
e
**Analysis and breakdown of results** for the two modeling approaches

**Fig. 1 | Data and model pipeline for this study.** **a**, Biodiversity and land-use data from 681 source studies comprising 25,987 sampling sites were joined with data on human population density, road network density, bioclimatic factors, and topography. **b**, We trained two statistical pressure-response models: 1) A mixed effects model with human pressures as fixed effects, where study-level random effects accounted for other variation in the data. 2) A hierarchical model with varying parameters at the biogeographic-taxonomic level, also including additional bioclimatic and topographic covariates. Geometric mean abundance was used to represent site-level alpha diversity, while beta diversity was calculated as the Bray-Curtis similarity between ecologically intact reference sites and other sites. **c**, Two cross-validation strategies were used: For interpolation ('standard' CV), folds were generated by splitting data at the site level, such that sites from a given study could be allocated to multiple folds. For extrapolation (cross-study validation), folds were based on splits at the study level, such that all sites from a given study only appeared in a single fold. **d**, Model accuracy was assessed through the correlation between predicted and observed values (Pearson's $r$) and the mean absolute error (MAE). **e**, The results were analyzed to understand performance differences and issues across model types, scenarios and countries. *Flag icons source: Freepik (CA, SE), iconset.co (US). Map source: The Transhumanist.*

global extent and broad taxonomic scope considered here. The amplified data limitations that come with this[13–15] makes evaluation more challenging[30] but nonetheless crucial[25]. To our knowledge, this has not been done for pressure-based biodiversity models with a global scope.

Here we quantitatively evaluate the generality of two different pressure-response models for projecting site-level biodiversity (overview in Fig. 1). Generality is operationalized as model accuracy when making spatially explicit, out-of-sample predictions in sampled ecological contexts (generalizability) as well as in other contexts (transferability). We leverage species inventory data from 25,987 sites in 96 countries, with broad taxonomic coverage. The data originates from 681 studies collated in the widely used PREDICTS database[31] (see Extended Data Fig. 1 and Extended Data Fig. 2). The first model uses a mixed effects structure[32] with human pressures as fixed effects, estimated on average across all sites. Random effects account for variation in data and responses between studies. This model structure is common in biodiversity meta-analysis[19] and is used for spatial projections in several indicators[22,24]. In the second model, we take a different approach and group data by biogeographic and taxonomic attributes, connected through a hierarchical model structure. Model parameters are estimated individually for sufficiently large groups, while sharing information across. Additionally, bioclimatic and topographic data complement the human pressure variables to create a richer model. The rationale is for the model to learn from more contextual information that can be used for out-of-sample predictions. Our aim is to explore the implications of biodiversity data availability and model design on generality, which can provide key insights for future refinement of large-scale biodiversity models.

# Results

### Distributions of alpha and beta diversity

We used geometric mean abundance (GMA)[33] to quantify site-level alpha diversity, reflecting the richness, abundance and evenness of a sampled species community. Data from each study were then normalized by the within-study maxima to obtain a common 0–1 scale across studies[30]. For beta diversity, we calculated the Bray-Curtis (BC)[34] compositional similarity between ecologically intact reference sites − consisting of minimally used primary vegetation − and sites with other land-use types within the same study[30]. This metric considers species–by–species presence and relative abundance between sites. The resulting GMA and BC distributions are shown in Extended Data Fig. 3. Both are clearly right-skewed, with an inflation of zeros caused by a mix of true absences and imperfect detection[35,36]. While GMA can be highest at any site in a study, the BC score is capped at reference site levels, which explains the lower mean and thinner tail of the latter distribution.

### Evaluation of model predictive accuracy

Human pressure variables for land-use[31], population density[37], and road network density[38] formed the core of all models (see variables in Extended Data Table 1). In the first model, these constituted the fixed effects, while study-level random intercepts and slopes accounted for inter-study differences in geographic and taxonomic scope, environment, and sampling. Spatial block intercepts further captured intra-study variation where available. The model structure and variable selection largely followed previous implementations[22,24,30,39], including the use of linear mixed models. In the second model, the hierarchical structure consisted of two levels: data was first grouped by biome and high-level taxonomic group, then subdivided by biogeographic realms to form 'regional biomes'[39] (see overview of hierarchical groups in Extended Data Fig. 4).

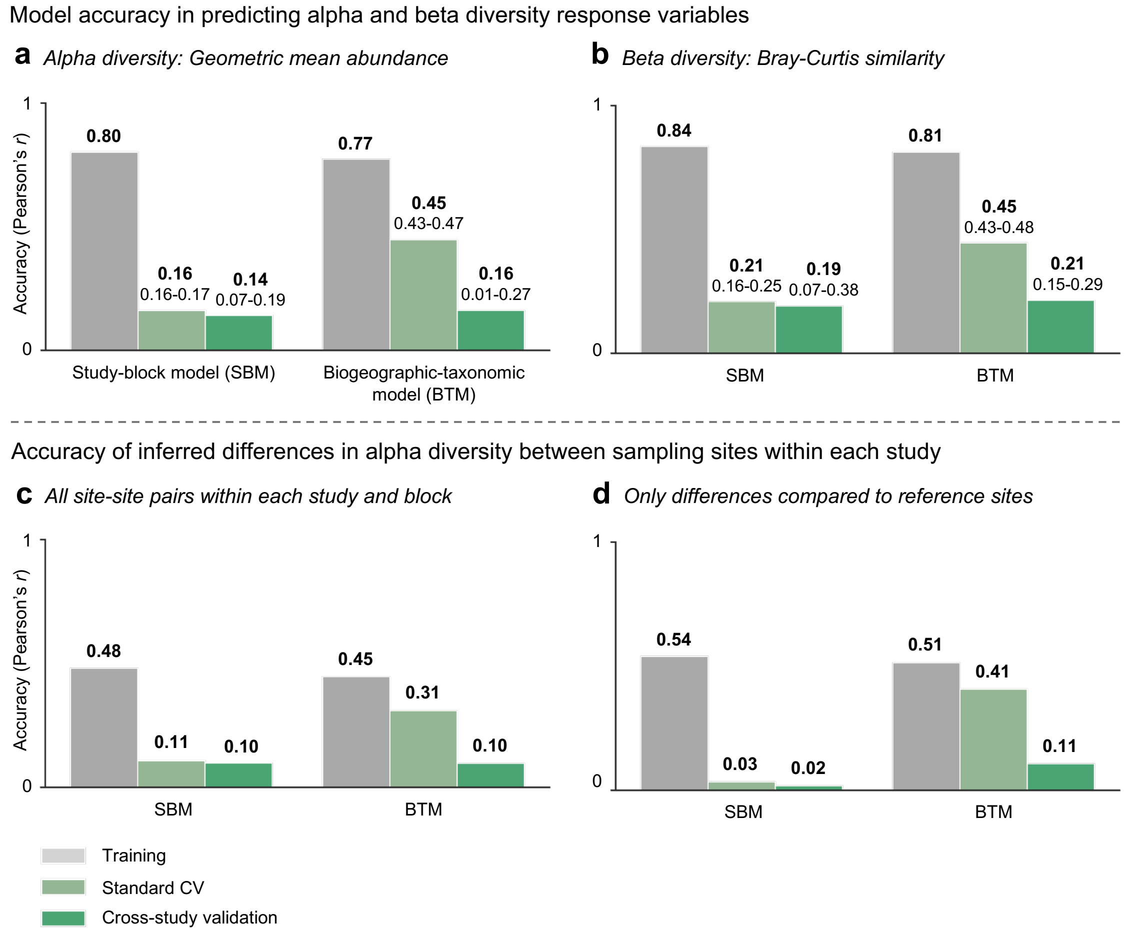


**Fig. 2 | Summary of model predictive accuracy. a–d**, Each panel shows the average predictive accuracy (Pearson's $r$ between predicted and observed values) of the study-block model (SBM) and biogeographic-taxonomic model (BTM). In-sample accuracy on training data is shown in grey, standard cross-validation (CV) accuracy in light green, and cross-study validation accuracy in dark green. For the CV results, the ranges indicate the minimum and maximum accuracies among folds. **a**, Alpha diversity, predictions of site-level geometric mean abundance. **b**, Beta diversity, predictions of Bray–Curtis similarity between pairs of ecologically intact reference sites and other sites, within each study. **c**, Differences in alpha diversity between all site–site pairs within every study and/or block, inferred from the alpha diversity models in (**a**). This was calculated from the predicted and observed deltas in alpha diversity between sites. **d**, Same as (**a**) but restricted to pairs containing an ecologically intact reference site.

Groups containing data from at least five studies used group-level parameters for predictions, while smaller groups were 'rolled up' to the level above to avoid overoptimistic interpolation results (from biome-taxa-realm to biome-taxa, or biome-taxa to the overall mean). Environmental variables included temperature and precipitation[40], and elevation and slope[41] (Extended Data Table 1). Study and block identifiers were included during training to avoid biased parameter estimates. For the implementation we chose a Bayesian hierarchical framework, to enable model regularization through weakly informative priors, handle overparameterization in small groups, and leverage statistical strength across groups[35,42,43].

A summary of model performance is shown in Fig. 2. In terms of alpha diversity (Fig. 2a), the study-block model (SBM, first model above) achieved predictive accuracies of 0.80 on training data, 0.16 in standard cross-validation (CV), and 0.14 in cross-study validation. This was measured as the Pearson correlation $r$ between observed and predicted values. The corresponding numbers for the biogeographic-taxonomic model (BTM) were 0.77 on training data, 0.45 in standard CV, and 0.16 in cross-study validation. For standard CV, folds were generated using stratified random sampling of *sites* across all studies, to assess generalizability (interpolation accuracy) within sampled contexts. Cross-study validation[44] evaluated transferability (extrapolation) to other contexts, where each non-overlapping fold contained a stratified random sample of *studies.* For beta diversity (Fig. 2b) the relative patterns were similar. The accuracies for the SBM were 0.84 (training), 0.21 (standard CV), and 0.19 (cross-study validation); for the BTM they were 0.81, 0.45, and 0.21, respectively. The beta diversity models included the same covariates as described above, plus differences in human population and road density, and the spatial and environmental distance, between sites[30]. Due to the large number of site pairs in the BC dataset a sub-sample was used, with inverse weighting based on study size to obtain a more balanced sample.

The results in Fig. 2a,b reveal notable gaps between in-sample and out-of-sample predictions. Three things stand out: i) The consistently low SBM accuracies, caused by reliance on a small set of highly averaged fixed effects. ii) The relatively higher BTM interpolation accuracies, although still low in absolute terms. This is explained by a combination of structure – a biogeographically and taxonomically explicit hierarchy with group-level parameters – and a richer covariate set. This is clearly beneficial when predicting alpha diversity for all observations globally, where relative distribution patterns depend on environmental factors, even when data have been normalized. It also helps for beta diversity, since compositional similarity within the context of a study is a function of both human pressures and natural species turnover. iii) Conversely, the equally low extrapolation performance of the BTMs compared to the SBMs, caused by distribution shifts in covariates. The drivers behind i) and iii) are further analyzed in the next section. It should be noted that the structural differences between the models imply some restrictions on a direct intercomparison. In the SBMs, the fixed effects used for prediction were estimated across all sites, covering the full spatial and taxonomic scope of the data. The site-level predictions, from applying those parameters to local human pressures, were then evaluated against site-level observed values that, naturally, represent a very small share of the full data scope. In the absence of taxonomically complete validation data, covering major biogeographic regions, this is still a useful approximation. The BTMs, on the other hand, were based on response variables that were split by broad taxonomic group. That makes evaluation of predictions specific to each biogeographic-taxonomic context, which is more interpretable. While this is a non-negligible difference, it does not prevent us from deriving valuable insights from the results.

Further, we looked at estimated and observed *differences* in alpha diversity *within* individual studies and spatial blocks (Fig. 2c,d), inferred from the predictions in Fig. 2a (note that these are not separately trained models). The more homogeneous context could provide a more 'level playing field' between the two model structures, in light of what we noted above. When evaluated over all site pairs (Fig. 2c), the SBM accuracies were at a somewhat lower level than the previous alpha predictions, with standard CV at 0.11 (vs 0.16) and cross-study validation at 0.10 (vs 0.14). The BTM also did worse here, for both standard CV (0.31 vs 0.45) and cross-study validation (0.20 vs 0.17). Finally, we repeated the analysis but restricted to only include pairs containing a reference site (Fig. 2d). The inferred site-site differences from the SBM were less accurate than those calculated for all site pairs, yet the opposite was true for the BTM model. It is not clear why that was the case, and since Fig. 2c,d are not based on a separately trained model, the results should be interpreted with some caution. At least it seems that the BTMs again benefited from the group-varying parameters, compared to the averaged SBM fixed effects.

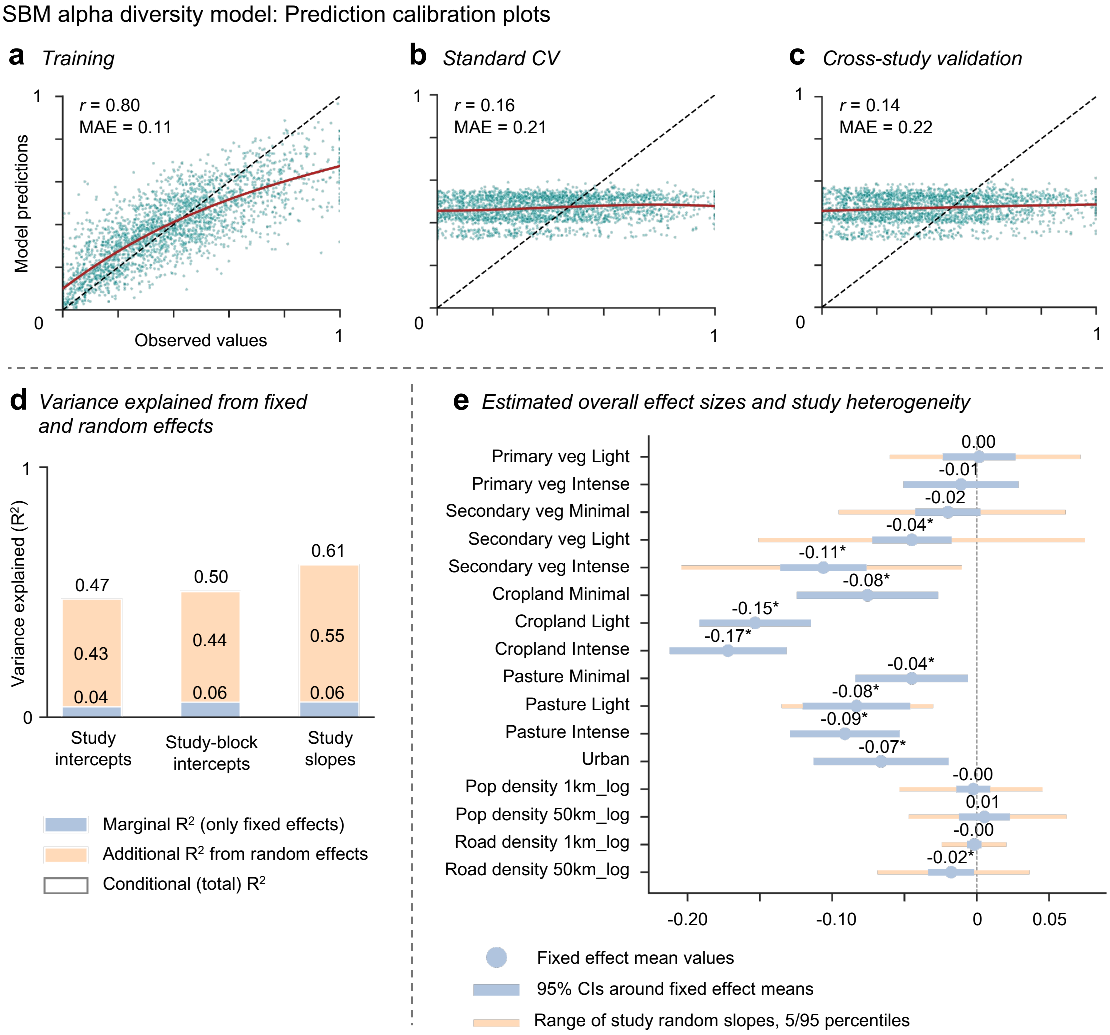


**Fig. 3 | SBM deep dive analysis. a–c**, Prediction calibration plots for the SBM alpha diversity model (showing a 10% random sample for visual clarity), for training data (**a**), standard CV (**b**) and cross-study validation (**c**). The black dashed lines indicate a perfect correlation, whereas the red solid lines show the best polynomial fit to the data. Areas where the solid line is above the dashed line indicate overbiased model predictions, and vice versa. **d**, Decomposition of variance explained between predictions using only the fixed effects (marginal $R^2$, light blue) and predictions including random effects (incremental contribution to conditional $R^2$, yellow). Total variance explained (conditional $R^2$) is shown on top of the bars. **e**, Estimated model parameters. Blue circles show the fixed effect means (across all studies), with blue bars representing 95% confidence intervals around the mean values. The yellow bars indicate study heterogeneity, the spread of study-level random effects (5–95th percentiles among all studies).

## Drivers of model generality gaps

The underlying patterns of the low SBM accuracies are clearly seen when comparing the in-sample predictions in Fig. 3a to the out-of-sample results in Fig. 3b,c. A key explanation behind this was the low attribution of variance explained[45] to the observable fixed effects (human pressures), relative to the random effects for studies and within-study spatial blocks (Fig. 3d). Since cross-validation simulates model performance on previously unseen sites, these learned random effects cannot be used to differentiate between observations when making out-of-sample predictions. In terms of the fixed effects, all land-use types except lightly used primary vegetation had,

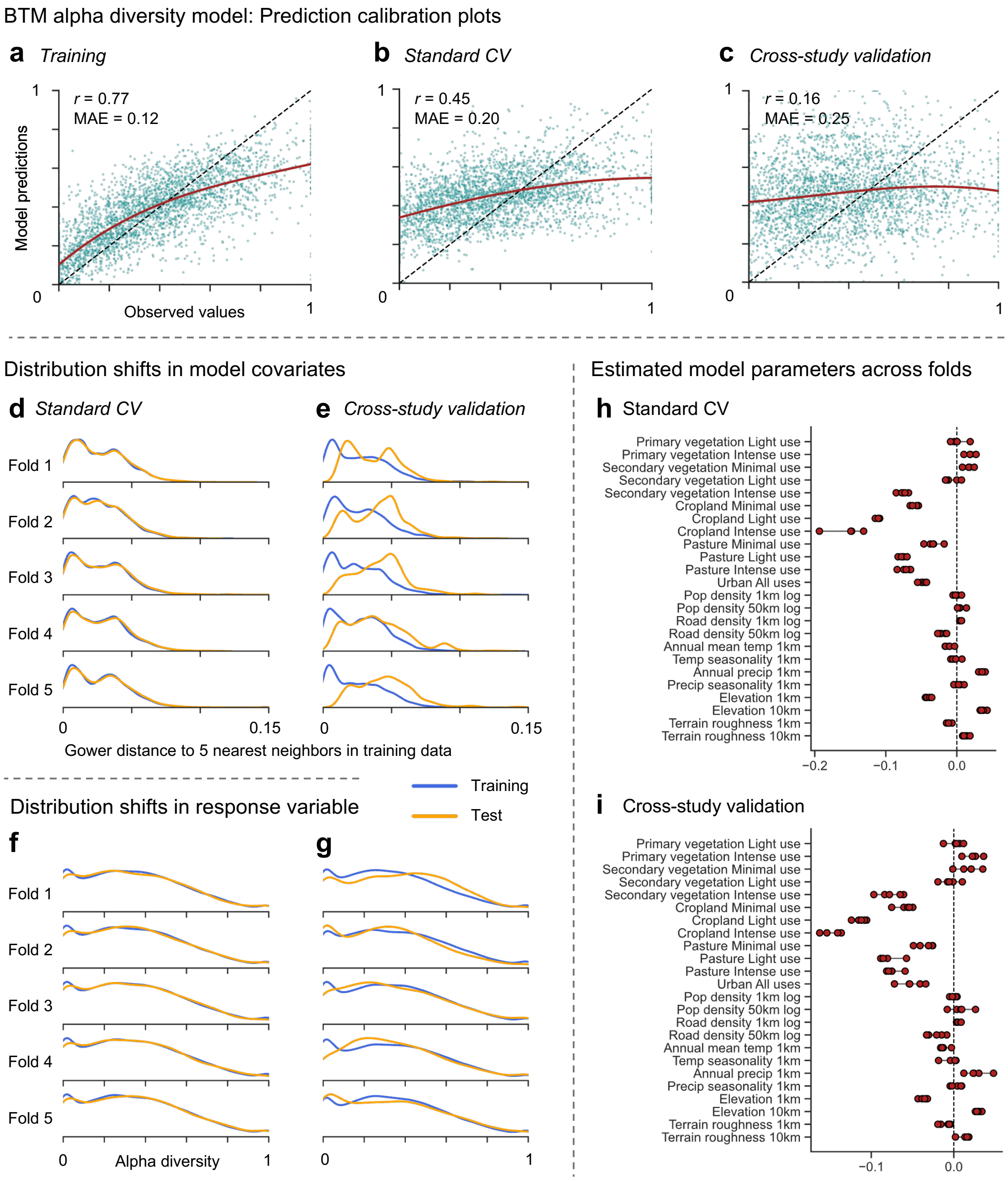


**Fig. 4 | BTM deep dive analysis. a–c**, Prediction calibration plots for the BTM alpha diversity model (corresponding to the SBM plots in Fig. 3). **d,e**, Distribution density plots obtained by calculating the multivariate Gower distance between the model covariates of each data point, and its five nearest neighbors in the training data (using a sample of 1,000 training points). This was done for all training data (blue) and test data (yellow). This gives an indication of how well the joint distribution of covariates align between training and test data in each cross-validation fold, for standard CV (**d**) and cross-study validation (**e**). **f,g**, Corresponding density plots of geometric mean abundance of the training and test data in each cross-validation fold, for standard CV (**f**) and cross-study validation (**g**). **h,i**, Posterior mean values of the model parameters at the highest hierarchical level, across all folds.

on average, negative estimated effects on diversity compared to the reference sites (Fig. 3e), several significant at the 5% level. However, there was also high cross-study heterogeneity around the mean effects, emphasizing the challenge of model generalization and transferability. The impacts of population and road density were ambiguous, possibly because these data are static in time.

The BTM interpolation results were relatively higher but still low in absolute terms. Fig. 4b shows clear calibration issues in the alpha diversity model: it overestimated predictions for small observed values, and vice versa for large ones. This is partially related to the high degree of skewness in the response variables (Extended Data Fig. 3). The poor extrapolation results (Fig. 4c) were caused by covariate distribution shifts between training and test folds, forcing the model to make out-of-distribution predictions[28,29]. This is clearly seen when comparing the train-test distribution alignment between the standard CV and cross-study validation runs (Fig. 4d,e). Here we calculated the multivariate Gower distance[46] between the covariates of each training and test point and its five nearest neighbors in the training data. A similar pattern, albeit less pronounced, could be seen for the response variable distributions (Fig. 4f,g). The distribution shifts translated into a larger spread in estimated parameters between folds, although less visible for the population-level parameters here (Fig. 4h,i). At the core, this has to do with model exposure to data. Standard CV involved sampling from a pool of 25,987 sites. For large samples, the model is exposed to a high proportion of the signal in the complete dataset in each training fold. In contrast, cross-study validation involved sampling studies from a heterogeneous pool of only 681. This led to spatial, environmental and taxonomic separation of training and test data. Interestingly, although the extrapolation accuracies were similar between the two models, the calibration plots (Figs. 3c and 4c) show very different patterns. The flexibility of the BTMs, an advantage for interpolation, here led to clear overfitting to the studies seen in training.

### Country-level average accuracies

The GBF-MF emphasizes country-level reporting of relevant biodiversity indicators[6,8]. As a complement to the previous analyses, we calculated the average accuracy of sampling sites located in each country. We used outputs from the BTM alpha and beta diversity models in the interpolation (standard CV) case. We only included countries with at least two studies and 25 sampling sites, to avoid outliers skewing the results. The analysis showed large variation in average accuracy across countries (Fig. 5). This reflects underlying differences in the predictive performance of different biogeographic-taxonomic groups, the level at which model parameters were estimated and used for predictions. In other words the heatmaps do not reflect overall expected accuracies per country, but the average of the subset of biogeographic regions, taxa, and sites represented in the training data. Data coverage varies substantially between geographies (see Extended Data Fig. 2), although this only partially explains model performance differences.

## Discussion

In this study, we quantitatively evaluated the predictive performance of two pressure-based biodiversity models, using a global and taxonomically broad dataset. Our results highlight clear limits to spatially explicit predictions across models (Fig. 2), related to data availability and model design. These findings suggest that the generality of large-scale biodiversity models remains a major challenge[19]. Some degree of generalization can be attained within contexts where there is enough data, but transferability to new contexts remains elusive. This aligns with prior results for models with smaller spatial and taxonomic scopes[27–29], even though the challenges are further amplified by scale here. Given the demand for global models and indicators[12], the systematic assessment presented here fills an important knowledge gap. Our results have important implications for the use and interpretation of pressure-based models and indicators.

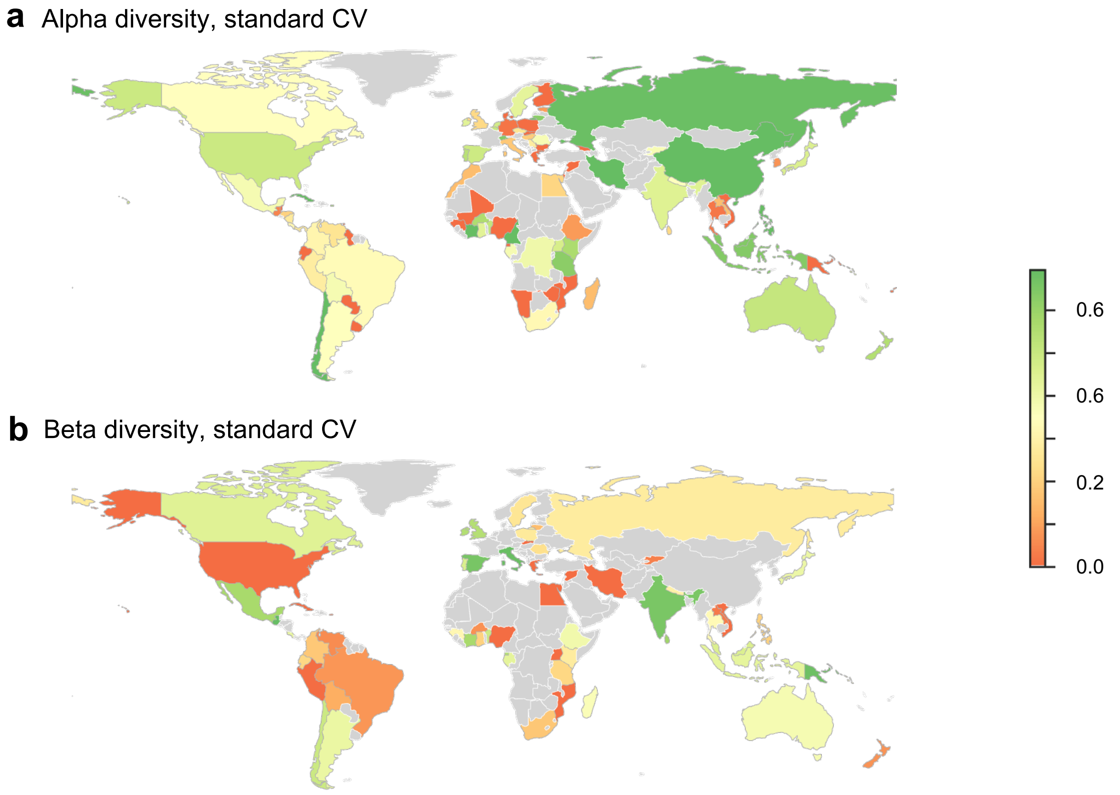


**Fig. 5 | BTM accuracy per country and biogeographic-taxonomic group. a,b**, Average model accuracy calculated per country for the BTM alpha (**a**) and beta (**b**) diversity models, when evaluated using standard CV. Only countries with at least two studies and 25 sampling sites are shown. Average accuracy values have been clipped at 0 and winsorized at the 5th and 95th percentiles to prevent outliers from skewing the heatmap gradient. Countries with insufficient data are shown in gray.

These insights can support future data collection and model development efforts, to further strengthen the critical role of biodiversity models for global policy and decision-making[16,18].

By comparing two structurally different models, we are able to highlight different aspects of the model generality challenge. Not surprisingly, models are only as good as the data they are trained on. Lingering data gaps and biases on a global level[13,14] puts fundamental constraints on the applicability of large-scale biodiversity models. This challenge is clearly shown by the decrease in predictive performance of the BTMs between interpolation (alpha diversity $r = 0.45$; beta $r = 0.45$) and extrapolation (alpha $r = 0.16$; beta $r = 0.21$) (Fig. 2a,b). The data distribution shifts (Fig. 4d,e) that forced the models to make out-of-distribution predictions (Fig. 4c) cannot be adequately addressed without additional data. More representative biodiversity data would imply fewer places where extrapolation is required. Another benefit is that it can enable interpolation within increasingly narrow contexts. For example, consider that the parameters in the BTMs were at the level of 'birds in moist broadleaf forests of the Neotropics'. That implies averaging over a great deal of underlying diversity in distributions and responses to pressures. The inter-country variation in the global heatmaps (Fig. 5) is an indication of this. However, even if we had used a more extensive dataset and achieved better results on average, we expect the patterns of varying interpolation accuracy and lack of extrapolation performance to persist. Our goal was to provide an illustrative comparison between two different modeling approaches, not exhaustively optimize model performance; that would also have involved a more extensive set of environmental and anthropogenic data.

Although data is foundational, the results also highlight the role of model design choices. The SBMs had consistently low performance across interpolation (alpha diversity $r = 0.16$; beta $r = 0.21$) and extrapolation (alpha $r = 0.14$; beta $r = 0.19$) (Fig. 2). The use of study-level random effects is sensible for estimating average effects of biodiversity drivers[32,47]. However, they also prevent effective learning, since the models attribute signal in the training data to variables that are discarded at the prediction stage (Fig. 3d). This shows that explanatory power is not always an indicator of predictive power[48,49]. In comparison, the explicit hierarchical structure and environmental covariates of the BTMs enabled some, albeit limited, generalization from sample to the broader context sampled (Figs. 2 and 4b). It is not surprising that a model trained using more contextually explicit information is better at predicting overall levels of alpha diversity (Fig. 2a). But even if the goal was mainly to differentiate between sites with different pressure levels, our results indicate that the BTMs have an edge through their varying-group parameters (Fig. 2b,c,d). Further, this hierarchical structure provides a dynamic mechanism for capturing the benefits of increasing data across different biogeographic-taxonomic contexts. Although a similar differentiation of effects could be achieved by training separate SBMs per group, it would be less dynamic and not leverage the full data to the same extent. Finally, it should be noted that despite their differences, the models estimated relatively consistent effects of land-use on alpha diversity (Figs. 3e and 4h), directionally in line with previous analyses[22,30,39]. This shows that land-use change is a strong driver of biodiversity loss[3,4]. Yet, the generally low predictive accuracies emphasize the challenging gap between effect size inference and spatially explicit predictions.

Scientifically robust indicators are essential for achieving the goals of the GBF[6,8]. Global implementations are appealing because they are readily applicable and offer promises of consistency[2]. But to effectively support decision-making, the limitations of model-based indicators must be transparent to users. One motivation behind this study was the use of pressure-based models for spatial projections[22,24] despite lack of systematic testing[25]. Our results show that the gaps between directional signals and precise estimates can be large. Maps where spatial resolution is not matched by predictive power can incur concrete risks, such as suboptimal conservation priorities or inaccurate sustainability reports. Countries with extensive data can overcome certain limitations by implementing their own models and indicators, in line with the GBF-MF[6,8]. However, that would still leave data-poor regions with a lack of accurate and actionable biodiversity insights on local scales[26]. Here, the global community must allocate resources to scale up the collection of standardized biodiversity data in prioritized areas and groups[13,50], leveraging cost-effective emerging technologies like environmental DNA, bioacoustics and camera traps. Similarly, companies that want to ensure accurate decision-making and reporting will often need to collect their own data[11], instead of only relying on off-the-shelf solutions.

Our study highlights some key research challenges for improving large-scale, predictive biodiversity models. In terms of data, there are clear opportunities to build dynamic and scalable pipelines that combine large quantities of sampling event data from repositories like GBIF, with the latest remote sensing data. Contributions of scientific study data to central repositories should be prioritized, where large-scale datsets are of particular value. More effort should also be invested into methodological development of top-down community models, like the ones used in this study. By aggregating biodiversity data first and then training models, they offer the necessary scalability for global applications. There are a lot of recent methodological developments in species distribution models that can be leveraged here. Methods that tackle data-sparse contexts and explore the limits of extrapolation are other interesting avenues of research. Finally, evaluating biodiversity models with such a large scope is a very complex task. Some key limitations to site-level evaluation and model intercomparison have been highlighted already. This also points to the need for dynamic evaluation frameworks that can handle scale – how do we know if a global model is good enough in a given ecological context or country?

In conclusion, models are essential for biodiversity monitoring, but our findings demonstrate that the predictive power of pressure-based models has clear limitations. Users need to be aware of these limitations, and the scientific community should prioritize the mobilization of data and development of refined models, to ensure that large-scale biodiversity indicators can reliably guide conservation efforts.

# Methods

## Data sources

*Biodiversity data:* All biodiversity data were obtained from the Projecting Responses of Ecological Diversity in Changing Terrestrial Systems (PREDICTS) project[31], a meta-database compiled from independent source studies and inventories. We combined the two publicly available releases of the database (from 2016 and 2022)[51,52] into one dataset. Before any filters (see further down), it contains data from 817 studies comprising 35,736 sampling sites in 101 countries, with 4,318,808 unique records across 53,925 species, collected between 1984 and 2018. The dataset mainly covers animals and plants, with the most common groups being insects and other arthropods, vascular plants, and vertebrate animals. Despite efforts to balance the data taxonomically, vertebrates are over-represented relative to their true diversity, while fungi and non-arthropod invertebrates are underrepresented. Geographically, the dataset has gaps in line with most biodiversity data sources; high-income regions like North America and Europe are well-sampled relative to their underlying biodiversity, tropical areas in Latin America have high coverage, while there are significant gaps in Africa and parts of Asia (see Extended Data Fig. 1 for an overview of PREDICTS data coverage). In 255 of the source studies, spatially adjacent sampling sites had been grouped into spatial blocks.

*Human pressure data:* The predominant land-use type and land-use intensity at each sampling site has been categorized by the PREDICTS team based on information in the source studies[31]. Land-use categories include primary vegetation, secondary vegetation (split into young, intermediate, mature, or indeterminate age), plantation forest, cropland, pasture and urban, while use intensity has been classified as minimal, light or intense. Data on human population density, expressed as the number of people per 1 km$^2$, came from the Gridded Population of the World, v4.11 (GPWv4.11) dataset[37], based on the 2010 round of Population and Housing Censuses (conducted 2005–2014) and adjusted to match the United Nations World Population Prospects country totals. The available data represents extrapolated numbers for the years 2000, 2005, 2010, 2015, and 2020. Information about road networks was taken from the Global Roads Open Access Data Set, v1 (gROADS)[38], a global layer of joined country road networks adjusted for topology. These data were collected between 1980 and 2010, with limited information on original road construction dates.

*Bioclimatic and topographic data:* Bioclimatic variables, such as annual averages, seasonality and extreme values of temperature and precipitation, were based on the WorldClim v2.1 dataset[40], frequently used in species distribution models and other biodiversity studies. The data represent average values over 1970–2000. Data on elevation, slope and terrain roughness came from the EarthEnv repository[41], constructed from global digital elevation models derived from satellite imagery and LiDAR measurements.

The spatial resolution of the land-use data depends on the specific sampling extent of a given site in each source study (only available for 13.1% of the sites). The resolution of the population density, bioclimatic and topographic variables is 30 arc-seconds (approximately 1 km$^2$ at the equator). The spatial accuracy of the road network data varies by country. Manual classification of land-use type and intensity could have introduced inconsistencies within and between studies, and because these data represent the predominant land-use at a site, it ignores potential mixed land-uses. The remote sensing and census data layers are all modeled to some extent, potentially introducing additional uncertainty and errors.

### Biodiversity metrics

We used two biodiversity metrics, the geometric mean abundance (GMA)[33], an alpha diversity metric, and the Bray-Curtis (BC) similarity[34], a beta diversity metrics. Combined, they provide insights into the relative level of community richness, abundance and compositional similarity between areas, and are suitable for detecting changes in biodiversity[33]. If we let $a_{si}$ represent the population abundance of species $s$ at sampling site $i$, and let $S_i$ represent the total number of species at the site, the GMA of that site can be calculated as

$$y_i = \exp\left(\frac{\sum_{s=1}^{S_i} \ln a_{si}}{S_i}\right).$$

The log transformation dampens the contribution of highly abundant (perhaps generalist or opportunistic) species. For a given total abundance, higher richness and evenness results in a greater GMA. For beta diversity, the BC similarity between two sampling sites $i$ and $j$ is given by

$$y_{ij} = \frac{2\sum_{s=1}^{S} \min(a_{si}, a_{sj})}{\sum_{s=1}^{S} \left(a_{si} + a_{sj}\right)},$$

where $S$ represents the total number of species found at both sites. This is equivalent to an abundance extension of the Sørensen index. The index takes a value of 0 if there are no shared species between sites $i$ and $j$, and a value of 1 if the species and their abundances are identical between the two sites (which of course is highly unlikely). Since the focus of this study is on models used to estimate biodiversity intactness, the BC similarity is only calculated between reference sites, consisting of minimally used primary vegetation sites, and all other sites within each study. This follows the approach in e.g. the BII[30].

One challenge with meta-databases is that the abundance distributions from different source studies will be greatly influence by taxonomic scope, biogeographic area and sampling method. Furthermore, while abundance is often expressed as a count of individuals, it can also be a proportion or density. The most viable approach is to normalize the GMA values within each study to a 0–1 scale, by dividing them by the within-study maxima (again following the BII)[30]:

$$y_i^{\text{norm}} = \frac{y_i}{\max(y_1, \ldots, y_I)}.$$

The BC index is already expressed on a 0–1 scale, so no further processing was needed. For simplicity, we let $y_i$ denote the normalized values from hereon. Although the sampling method is consistent between sampling sites within a given source study, sampling effort between sites can vary, so effort-adjusted abundance numbers (already provided in the PREDICTS data) were used in all analyses. Since both diversity metrics require abundance data, we filtered out studies that only recorded presences and absences. To mitigate the impact of extreme abundance values, we identified outlier locations using the interquartile range (IQR) method and removed site-level observations where $y_i > 1.5\,\mathrm{IQR}$, within each study.

The distributions of the alpha (GMA) and beta (BC similarity) metrics are shown in Extended Data Fig. 3. Both distributions are non-symmetric and skewed to the right. The shapes are similar, but the GMA distribution has a greater mean than the BC index, since the BC formula reflects natural species community turnover across the landscape. There is an inflation of zeros in both distributions, and an inflation of ones in the GMA data. A site-level abundance of zero can either be the result of true absence of the surveyed species, or the result of imperfect detection, a well-known issue in biodiversity data. In the best case, noise that arises from imperfect detection is randomly distributed across studies and sites, but we acknowledge that there could be more systematic detection biases in the data. However, there is no feasible way to construct a separate detection probability model[35,36] for such a heterogeneous dataset, based on the data that we have available. The inflation of zeros in the BC distribution is a mix of observed zero abundances and complete dissimilarity in species composition between sites. The inflation of ones in the GMA data is an artifact of the normalization procedure described earlier, since every study will contribute a one (its maximum abundance site) to the overall data pool.

In the main results, we used data from all studies except for what was subject to the filters described above (such as requiring abundance data). It should be noted, however, that one issue is that many source studies contain very few sites: 18.8% contain 5 sites or less, and 60.9% contain 25 or fewer (see Extended Data Fig. 5). There is a risk that such small studies contribute more noise than signal to the overall pool of data used to train the models. Informally, this is because there are not enough observations from the context of a given study to reliably relate its species observations to different human pressures and environmental conditions. Through inspection of study-level histograms, we found that at least 25–50 sites would be required to produce reasonable, somewhat continuous distributions of data at the study-level.

The BTMs used taxonomic groups as part of the hierarchical model structure (see Extended Data Fig. 1), and since some larger studies sampled species across several such taxonomic groups, sites were consequently split into multiple observations. The BTM alpha model had 31,242 observations in total, in comparison to the 25,987 sites in scope, which also equaled the number of observations in the SBM model. Since there were 426,573 site pairs in the BC dataset after the standard filters, we used sub-sampling to obtain a more reasonably-sized dataset for model training and evaluation. As the contribution of large studies with many reference sites became disproportional when generating pair-wise comparisons, the sampling scheme aimed to maintain relative balance between studies. Studies contributed site-pairs linearly up to a threshold of 300, after which the contribution was weighted by the square root of the potential site-pairs. There was also a cap of 3,000 pairs per study.

### Data likelihood functions

It is clear from Extended Data Fig. 3 that both GMA and BC distributions are non-symmetric and bounded between 0 and 1, with inflation of zeros and some inflation of ones for the GMA. This suggests that a zero-inflated or zero-one-inflated beta distribution would be the most appropriate likelihood function to describe the data. However, since our goal was to predict biodiversity on highly aggregated levels, it seemed counterproductive to explicitly model the zero-inflation, as true absences of entire species groups in a given area are unlikely. Further, the inflated ones are scaling artifacts that did not warrant explicit modeling either. During initial testing of the BTMs, we compared the results of a regular beta distribution against a Gaussian likelihood, the latter representing a simplified assumption about the data generating process. Although the shapes of the beta posterior distributions were more similar to the observed data, compared to its Gaussian counterpart, the predictive accuracy of the Gaussian models were generally higher. Since the aim of the study was to benchmark predictive performance, and Gaussian models are computationally simpler, with more interpretable model coefficients, we decided to use Gaussian likelihood functions across all models.

### Model covariates

Variables based on land-use[31], population density[37], and road network density[38] data, formed the core part of all models in the study. For land-use, we combined type and intensity into a set of categorical variables; records where this information was unknown were filtered out. For consistency with global land-use layers, minimally used plantation forest was grouped with lightly used secondary forest and other plantation forest (light and intense use) was grouped with intensely used secondary forest[30]. The categorical land-use variables were one-hot encoded, with the model intercepts representing minimally used primary vegetation reference sites. Human population density and road network density at 1 and 50 km$^2$ scales, log transformed to reduce skewness, were also part of all models. The BHMs additionally included the following bioclimatic and topographic variables: annual mean precipitation and temperature (1 km$^2$), temperature and precipitation seasonality (1 km$^2$), elevation (1 and 10 km$^2$), and terrain roughness (1 and 10 km$^2$). See full list of model covariates in Extended Data Table 1.

To derive the continuous variables, the coordinates of each sampling site were first projected from global EPSG:4326 to local UTM format, before generating circular polygons of different spatial extents. The equal-area projections ensure that the calculated values are comparable across all locations, regardless of distance to the equator. For each raster dataset (human population density, bioclimatic, topographic) the mean value of each polygon was calculated (including partially covered pixels), after which the polygons were reprojected to the global format. For the road network data, a similar approach was used to derive road density as the combined length of all roads within each polygon. Population density data were interpolated between the available years in the GPW dataset, and back to the earliest PREDICTS data (1984), assuming an exponential growth rate[53]. The population densities were matched to the sampling year of each site; the other covariates are static layers.

For the beta diversity (BC) models, we used the alpha model covariates as a starting point. Here, the land-use variables described the conditions at the non-reference site in each pair, since the reference sites constitute the model intercepts (minimally used primary vegetation). Differences in the human population and road density variables were calculated for each site pair. Additionally, the spatial (Haversine) distance, and multivariate environmental (Gower) distance[46], were included to control for natural compositional turnover[30]. The Haversine distance was normalized by the median sampling extent among all studies. The Gower distance was calculated using the following variables, all at the 1 km$^2$ scale: maximum and minimum temperature of the warmest and coldest months, respectively, precipitation of the wettest and driest months, and elevation. All continuous variables were standardized prior to model training, and evaluated for collinearity within their respective category; no correlation coefficients exceeded 0.5.

### SBMs: Linear mixed models

The SBM model structure consisted of fixed effects, parameters that represent average values across all studies, and random effects, which describe group-level variation around the fixed effects[32,47]. The study-level intercepts and slopes accounted for inter-study differences in geographic and taxonomic scope, environment, and sampling. For studies that grouped spatially adjacent sites into blocks, corresponding intercepts were included to capture intra-study differences. Since all site-level species observations were aggregated to form the response variables, the fixed effects represented average effect sizes across all geographies and taxa[22,23,30]. We let $\boldsymbol{\beta}$ denote the fixed effects and let $\boldsymbol{\gamma}_s$ denote the study-level random effects (both including intercept terms), for the study $s$ that site $i$ belongs to. The block intercepts are denoted by $\gamma_b$ for each block $b$ within study $s$. Hence, the different effects constitute a hierarchical structure, from the population of all studies to individual source studies, and from studies to blocks. For GMA, the site-level regression model can then be written as

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{x}_i^\top \boldsymbol{\gamma}_s + \gamma_b + \epsilon_i.$$

We fitted a full set of random slopes at the study level, that is, the random effect covariates are the same as the fixed effect ones, $\mathbf{x}_i$. The $\boldsymbol{\gamma}_s$ parameters are assumed to be uncorrelated and vary around the fixed effects with mean zero and variance $\sigma_\gamma^2$, such that $\boldsymbol{\gamma}_s \sim \mathcal{N}(0, \sigma_\gamma^2)$. The errors $\epsilon_i$, which describe remaining within-study variance, are assumed to be independent within and between studies, with fixed variance $\sigma^2$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. In the beta diversity model, $y_{ij}$ denotes the BC similarity between some site $i$ and a reference site $j$. In addition to the covariates $\mathbf{x}_i$ for the non-reference site, we let $\mathbf{z}_{ij}$ denote the set of delta measures calculated between the two sites: the difference in the population and road density, the spatial distance, and the environmental distance. If we let $\boldsymbol{\beta}^\Delta$ and $\boldsymbol{\gamma}_s^\Delta$ denote the parameter vectors associated with the difference terms, we can write the beta model as

$$y_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\beta}^\Delta + \mathbf{x}_i^\top \boldsymbol{\gamma}_s + \mathbf{z}_{ij}^\top \boldsymbol{\gamma}_s^\Delta + \gamma_b + \epsilon_i.$$

The SBM linear mixed models were implemented in `pymer4`[54], a Python wrapper around the `lme4` package in R, and fitted using restricted maximum likelihood.

**BTMs: Bayesian hierarchical models**

The SBM hierarchical structure is based on how data was collected, through studies that sampled sites, sometimes organized into spatial blocks. In the BTMs, we instead built the hierarchical structure around the biome where a site is located and the broad taxonomic group(s) sampled at that site. Biomes are distinct in their environmental conditions and underlying biodiversity patterns, suggesting that the effect sizes of both natural and anthropogenic drivers will be different between them. The taxonomic grouping was based on differences in spatial distribution patterns[55] and potentially differentiated responses to drivers, shown in Extended Data Fig. 1. At the same time, data limitations and taxonomic imbalances implied that most taxa were aggregated at very high levels. Given the geographic and taxonomic extent of the data, we did not want to make assumptions about any hierarchical relationships between biomes and taxonomic groups. We therefore combined both attributes at the same level, such that the models only had one hierarchical level below the overall population, which represented the average effects across all data.

Although this model structure could have been implemented as a frequentist linear mixed model, the Bayesian formulation enabled additional flexibility, stronger borrowing of statistical strength across groups, and regularization through weakly informative priors[35,42,56]. In the Bayesian models, groups with low information content will be pulled towards the population mean, while groups with sufficient statistical power are able deviate from those mean values. Still, it should be noted that the main comparison in our study is between the overall model approach and structure, not the frequentist versus Bayesian implementation. We let $b = 1, \ldots, B$ denote biomes and $t = 1, \ldots, T$ the taxonomic groups, with model intercepts denoted by $\alpha$ and other model parameters by $\beta$. To make observations conditionally independent during training, we also included study and spatial block identifiers $\gamma_{sb}$ as control variables. These were used in model training but not for any of the out-of-sample predictions. The alpha diversity (GMA) model for a site $i$ in hierarchical group $b, t$ can be written as

$$y_i = \alpha_{bt} + \mathbf{x}_i^\top \boldsymbol{\beta}_{bt} + \gamma_{sb} + \epsilon_i.$$

where all the parameters are allowed to vary at the group level. The corresponding beta diversity model can be expressed as

$$y_i = \alpha_{bt} + \mathbf{x}_i^\top \boldsymbol{\beta}_{bt} + \mathbf{z}_{ij}^\top \boldsymbol{\beta}_{bt}^\Delta + \gamma_{sb} + \epsilon_i.$$

Generally, the priors on the model parameters were assumed to be normal and uncorrelated, analoguous to the SBM case. Since variances are strictly positive, the variance priors were assumed to be half-normal. At the population-level we used the following hyperpriors:

$$\begin{aligned} \alpha: \ \mu_\alpha &\sim \mathcal{N}(0.5, 0.1^2), \ \sigma_\alpha \sim \text{Half-Normal}(0.1^2), \\ \beta: \ \mu_\beta &\sim \mathcal{N}(0, 0.1^2), \ \sigma_\beta \sim \text{Half-Normal}(0.1^2). \end{aligned}$$

The intercept hyperprior mean of 0.5 was simply the midpoint of the response variable distributions. The standard deviation terms of 0.1 were chosen based on prior predictive checks, in order to constrain the prior predictive distribution to a reasonable range, given the 0-1 bounded data, and achieve a high degree of regularization of the model. Based on these hyperpriors, the priors on the parameters of each biome-taxa group were given by

$$\alpha_{bt} \sim \mathcal{N}(\mu_\alpha, (\sqrt{n} - 1) \cdot \sigma_\alpha^2), \ \beta_{bt}^{(k)} \sim \mathcal{N}(\mu_\beta, (\sqrt{n} - 1) \cdot \sigma_\beta^2) \text{ for } k = 1, \ldots, K,$$

where the $K$ is the number of regression parameters and $(\sqrt{n} - 1)$ is an adaptive shrinkage factor to further constrain the prior variance of hierarchical groups containing few studies, such that those groups are pulled more strongly towards the population means. Since many groups contain quite few studies (Extended Data Fig. 1), there is a risk that the estimated group-level parameters become specific to a few studies, rather than generally applicable to a biome-taxonomic group at large. To further prevent overoptimistic interpolation accuracies, groups with less than five studies were 'rolled up' used population-level mean parameters for their predictions.

The BTMs were implemented in `PyMC`[57], using the No-U-Turn Sampler (NUTS) with a `numpyro` backend. For all experiments, we ran four parallel sampling chains for stability, using 1,000 tuning samples that were discarded and 500 posterior draws. A high target acceptance rate of 0.95 was used to avoid divergences when sampling from the complex posterior distribution.

### Cross-validation and performance metrics

We evaluated model generality and performance using two complementary cross-validation (CV) approaches. Five CV folds were used in all validation runs, with stratified sampling to ensure representative folds. The stratification was based on biomes and realms for the SBMs, and on biomes and taxonomic groups for the BTMs, reflecting how site-level species observations were aggregated in each model. The first approach, using 'standard' CV, assessed how well models could make predictions in environmental and taxonomic contexts similar to the data they were trained on. This evaluated model generalizability from sample to sampled population, also referred to as interpolation accuracy. The folds were generated by sampling at the site level among all studies within a given stratum. This implies that sites from a given study were split among several folds, creating overlap in studies, but not sites, between folds.

The second approach, based on cross-study validation[44], was used to assess how well the models could make predictions in new contexts; specifically, whether model learning is transferable between different source studies. In this case, folds were constructed by sampling at the study level, such that all sites from a given study ended up in a single fold. The only exception was if a study spanned several strata (e.g. biomes), in which case it could appear in more than one fold. This approach led to a clear spatial and environmental separation of training and test data, while also reflecting taxonomic and methodological differences between studies. Although the numbers of studies per fold were roughly equal, one consequence of the cross-study validation procedure was that folds in some iterations became unbalanced in terms of the number of sites, due to the large spread in the size of different studies.

We defined model accuracy as the Pearson correlation coefficient between predicted values $\widehat{\mathbf{y}}$ and observed values $\mathbf{y}$:

$$r(\widehat{\mathbf{y}}, \mathbf{y}) = \frac{\mathrm{Cov}(\widehat{\mathbf{y}},\, \mathbf{y})}{\sqrt{\mathrm{Var}(\widehat{\mathbf{y}})}\,\sqrt{\mathrm{Var}(\mathbf{y})}}.$$

We chose the correlation coefficient since it is widely recognized, easy to interpret, and provides a normalized measure of performance (values lie between $-1$ and 1). It directly relates to the calibration plots in Figs. 3 and 4; a well-calibrated model should exhibit a linear relationship between predictions and observations. As a complementary metric, we used the mean absolute error (MAE), defined as

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y}_i| \,.$$

Since all the data are on a 0–1 scale, this metric can be interpreted as the average percentage point deviation between predicted and observed values.

# References

1. Leclère, D. *et al.* Bending the Curve of Terrestrial Biodiversity Needs an Integrated Strategy. *Nature* **585,** 551–556 (2020).
2. IPBES. *The Global Assessment Report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, Bonn, 2019).
3. Keck, F. *et al.* The Global Human Impact on Biodiversity. *Nature* **641,** 395–400 (2025).
4. Jaureguiberry, P. *et al.* The Direct Drivers of Recent Global Anthropogenic Biodiversity Loss. *Sci. Adv.* **8,** eabm9982 (2022).
5. Cardinale, B. J. *et al.* Biodiversity Loss and Its Impact on Humanity. *Nature* **486,** 59–67 (2012).
6. CBD. *Monitoring Framework for the Kunming-Montreal Global Biodiversity Framework* (Convention on Biological Diversity, Montreal, 2025).
7. CBD. *Kunming-Montreal Global Biodiversity Framework* (Convention on Biological Diversity, Montreal, 2022).
8. Affinito, F., Williams, J. M., Campbell, J. E., Londono, M. C. & Gonzalez, A. Progress in Developing and Operationalizing the Monitoring Framework of the Global Biodiversity Framework. *Nat. Ecol. Evol.* **8,** 2163–2171 (2024).
9. SBTN. *Science-Based Targets for Nature: Initial Guidance for Business* (Science Based Target Network, 2020).
10. TNFD. *Recommendations of the Taskforce on Nature-related Financial Disclosures* (Taskforce on Nature-related Financial Disclosures, London, 2023).
11. Initiative, N. P. *Draft State of Nature Metrics for Piloting* (Nature Positive Initiative, 2025).
12. Burgess, N. D. *et al.* Global Metrics for Terrestrial Biodiversity. *Annu. Rev. Environ. Resour.* **49,** 673–709 (2024).
13. Hughes, A. C. *et al.* Sampling Biases Shape Our View of the Natural World. *Ecography* **44,** 1259–1269 (2021).
14. Chapman, M. *et al.* Biodiversity Monitoring for a Just Planetary Future. *Science* **383,** 34–36 (2024).
15. Boyd, R. J., Powney, G. D. & Pescott, O. L. We Need to Talk about Nonprobability Samples. *Trends Ecol. Evol.* **38,** 521–531 (2023).
16. Purvis, A. Bending the Curve of Biodiversity Loss Requires a 'Satnav' for Nature. *Phil. Trans. R. Soc. B* **380,** 20230210 (2025).
17. Damania, R. *et al.* *Nature's Frontiers: Achieving Sustainability, Efficiency, and Prosperity with Natural Capital* (World Bank, Washington, DC, 2023).
18. Zurell, D. *et al.* Predicting the Way Forward for the Global Biodiversity Framework. *Proc. Natl. Acad. Sci. USA* **122,** e2501695122 (2025).
19. Spake, R. *et al.* Improving Quantitative Synthesis to Achieve Generality in Ecology. *Nat. Ecol. Evol.* **6,** 1818–1828 (2022).
20. Maris, V. *et al.* Prediction in Ecology: Promises, Obstacles and Clarifications. *Oikos* **127,** 171–183 (2018).
21. Scholes, R. J. & Biggs, R. A Biodiversity Intactness Index. *Nature* **434,** 45–49 (2005).
22. Newbold, T. *et al.* Global Effects of Land Use on Local Terrestrial Biodiversity. *Nature* **520,** 45–50 (2015).
23. Newbold, T. *et al.* Has Land Use Pushed Terrestrial Biodiversity beyond the Planetary Boundary? A Global Assessment. *Science* **353,** 288–291 (2016).

24. Schipper, A. M. *et al.* Projecting Terrestrial Biodiversity Intactness with GLOBIO 4. *Glob. Change Biol.* **26,** 760–771 (2020).
25. Martin, P. A., Green, R. E. & Balmford, A. The Biodiversity Intactness Index May Underestimate Losses. *Nat. Ecol. Evol.* **3,** 862–863 (2019).
26. Clements, H. S. *et al.* A Place-Based Assessment of Biodiversity Intactness in Sub-Saharan Africa. *Nature,* 1–9 (2025).
27. Roberts, D. R. *et al.* Cross-validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* **40,** 913–929 (2017).
28. Meyer, H. & Pebesma, E. Predicting into Unknown Space? Estimating the Area of Applicability of Spatial Prediction Models. *Methods. Ecol. Evol.* **12,** 1620–1633 (2021).
29. Meyer, H. & Pebesma, E. Machine Learning-Based Global Maps of Ecological Variables and the Challenge of Assessing Them. *Nat. Commun.* **13** (2022).
30. De Palma, A. *et al.* Annual Changes in the Biodiversity Intactness Index in Tropical and Subtropical Forest Biomes, 2001–2012. *Sci. Rep.* **11,** 20249 (2021).
31. Hudson, L. N. *et al.* The Database of the PREDICTS (Projecting Responses of Ecological Diversity In Changing Terrestrial Systems) Project. *Ecol. Evol.* **7,** 145–188 (2017).
32. Harrison, X. A. *et al.* A Brief Introduction to Mixed Effects Modelling and Multi-Model Inference in Ecology. *PeerJ* **6,** e4794 (2018).
33. Santini, L. *et al.* Assessing the Suitability of Diversity Metrics to Detect Biodiversity Change. *Biol. Conserv.* **213,** 341–350 (2017).
34. Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* **27,** 325–349 (1957).
35. Dorazio, R. M. Bayesian Data Analysis in Population Ecology: Motivations, Methods, and Benefits. *Popul. Ecol.* **58,** 31–44 (2016).
36. Wu, G., Holan, S. H., Nilon, C. H. & Wikle, C. K. Bayesian Binomial Mixture Models for Estimating Abundance in Ecological Monitoring Studies. *Ann. Appl. Stat.* **9,** 1–26 (2015).
37. Center for International Earth Science Information Network-CIESIN-Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11. NASA Socioeconomic Data and Applications Center (SEDAC) https://doi.org/10.7927/H49C6VHW (2017).
38. Center for International Earth Science Information Network-CIESIN-Columbia University. Global Roads Open Access Data Set, Version 1 (gROADSv1). NASA Socioeconomic Data and Applications Center (SEDAC) https://doi.org/10.7927/H4VD6WCT (2013).
39. Bevan, P. A. *et al.* Regional Biomes Outperform Broader Spatial Units in Capturing Biodiversity Responses to Land-use Change. *Ecography* **2025,** e07318 (2024).
40. Fick, S. E. & Hijmans, R. J. WorldClim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas. *Int. J. Climatol.* **37,** 4302–4315 (2017).
41. Amatulli, G. *et al.* A Suite of Global, Cross-Scale Topographic Variables for Environmental and Biodiversity Modeling. *Sci. Data* **5,** 180040 (2018).
42. Lemoine, N. P. Moving beyond Noninformative Priors: Why and How to Choose Weakly Informative Priors in Bayesian Analyses. *Oikos* **128,** 912–928 (2019).
43. Gelman, A. Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics* **48,** 432–435 (2006).
44. Bernau, C. *et al.* Cross-Study Validation for the Assessment of Prediction Algorithms. *Bioinformatics* **30,** i105–i112 (2014).
45. Nakagawa, S. & Schielzeth, H. A General and Simple Method for Obtaining $R^2$ from Generalized Linear Mixed-effects Models. *Methods. Ecol. Evol.* **4,** 133–142 (2013).

46. Gower, J. C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **27,** 857–871 (1971).

47. Pinheiro, J. C. & Bates, D. M. *Mixed-Effects Models in S and S-PLUS* Ch. 1 (Springer, 2000).

48. Shmueli, G. To Explain or to Predict? *Stat. Sci.* **25,** 289–310 (2010).

49. Tredennick, A. T., Hooker, G., Ellner, S. P. & Adler, P. B. A Practical Guide to Selecting Models for Exploration, Inference, and Prediction in Ecology. *Ecology* **102,** e03336 (2021).

50. Gonzalez, A. *et al.* A Global Biodiversity Observing System to Unite Monitoring and Guide Action. *Nat. Ecol. Evol.* **7,** 1947–1952 (2023).

51. Hudson, L. *et al.* The 2016 Release of the PREDICTS Database V1.1. Natural History Museum https://doi.org/10.5519/J4SH7E0W (2023).

52. Contu, S. *et al.* Release of Data Added to the PREDICTS Database. Natural History Museum https://doi.org/10.5519/JG7I52DG (2022).

53. Goldewijk, K. K. Three Centuries of Global Population Growth: A Spatial Referenced Population (Density) Database for 1700–2000. *Popul. Environ.* **26,** 343–367 (2005).

54. Jolly, E. Pymer4: Connecting R and Python for Linear Mixed Modeling. *J. Open Source Softw.* **3,** 862 (2018).

55. Jenkins, C. N., Van Houtan, K. S., Pimm, S. L. & Sexton, J. O. US Protected Lands Mismatch Biodiversity Priorities. *Proc. Natl. Acad. Sci. USA* **112,** 5081–5086 (2015).

56. Gelman, A., Hill, J. & Yajima, M. Why We (Usually) Don't Have to Worry About Multiple Comparisons. *J. Res. Educ. Eff.* **5,** 189–211 (2012).

57. Abril-Pla, O. *et al.* PyMC: A Modern, and Comprehensive Probabilistic Programming Framework in Python. *PeerJ Comput. Sci.* **9,** e1516 (2023).

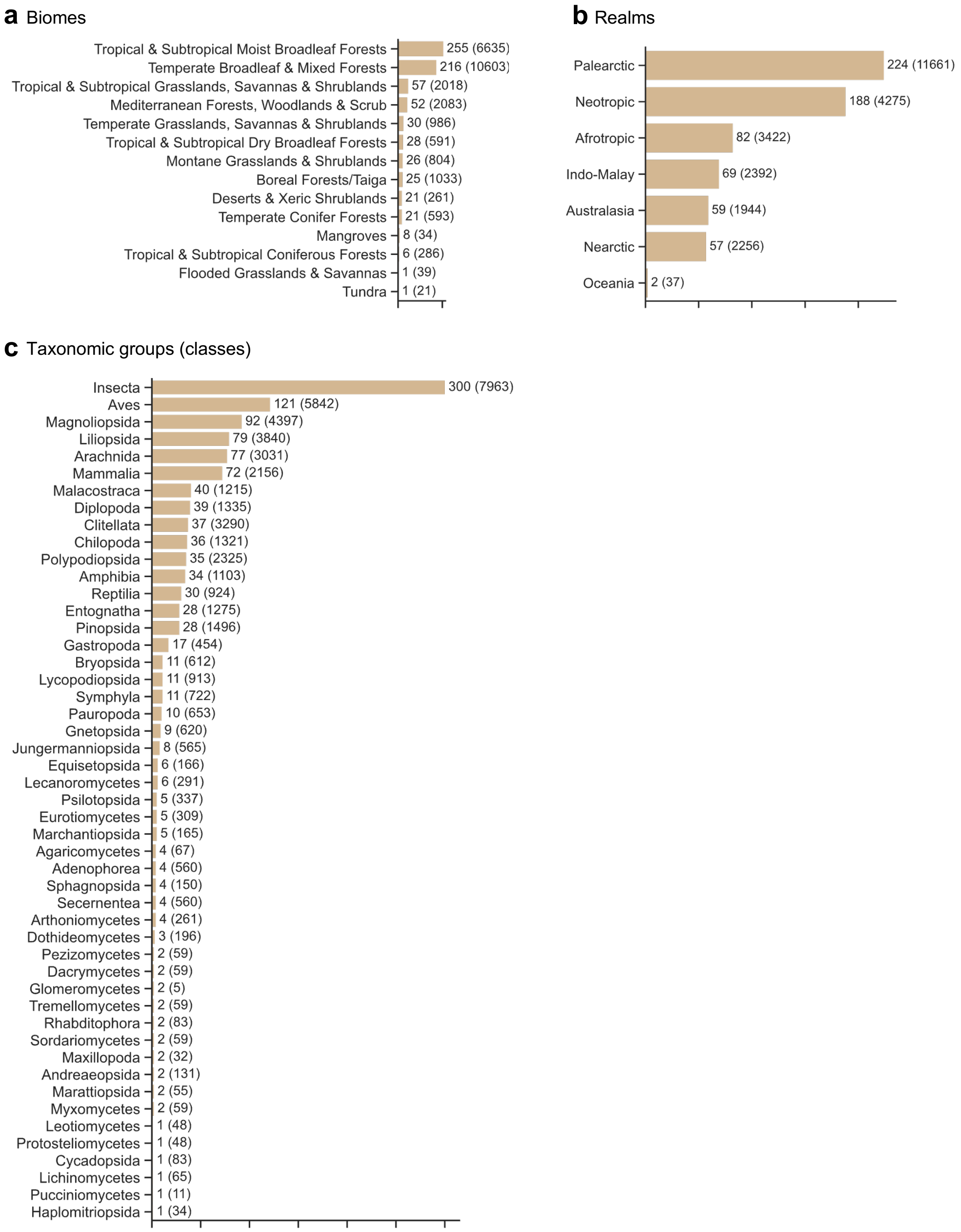
## Author contributions

JN conceptualized the study and developed the overall approach with support from JRS, LM and TA. JN processed the data, implemented the models, and conducted the statistical analyses. All authors provided continuous feedback on the results. JN wrote the first draft of the paper. All authors edited and approved the final version.

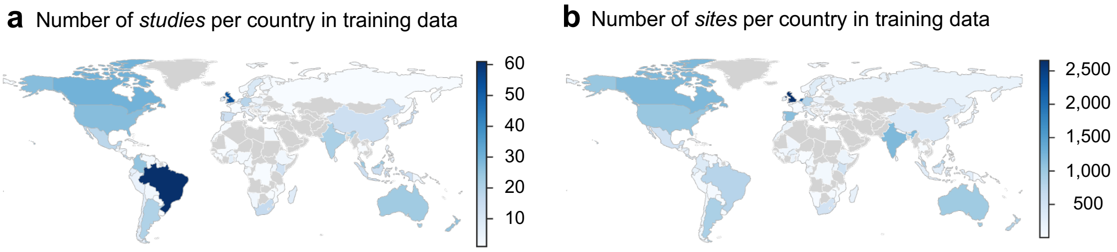
## Competing interests

The authors declare no competing interests.

# Extended Data



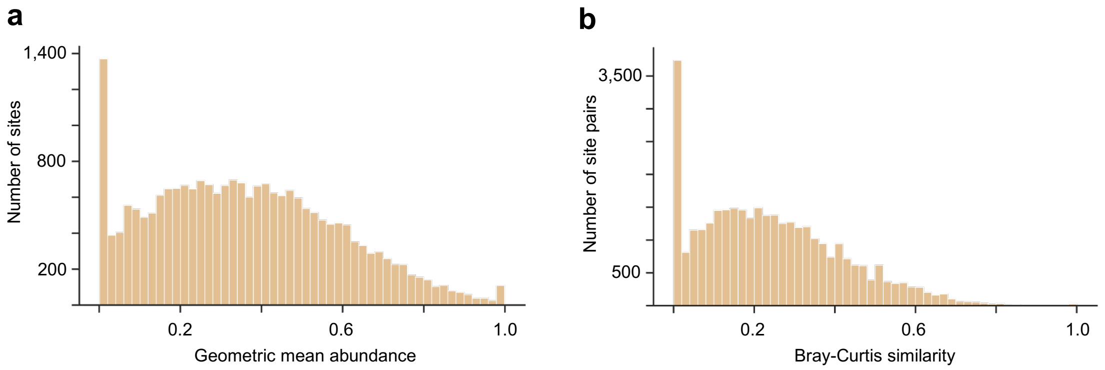
**Extended Data Fig. 1 | PREDICTS data coverage. a-c**, Number of source studies and sampling sites (in parenthesis) in the model data, split by biomes (**a**), realms (**b**), and taxonomic classes (**c**).


a Number of *studies* per country in training data
60
50
40
30
20
10
b Number of *sites* per country in training data
2,500
2,000
1,500
1,000
500

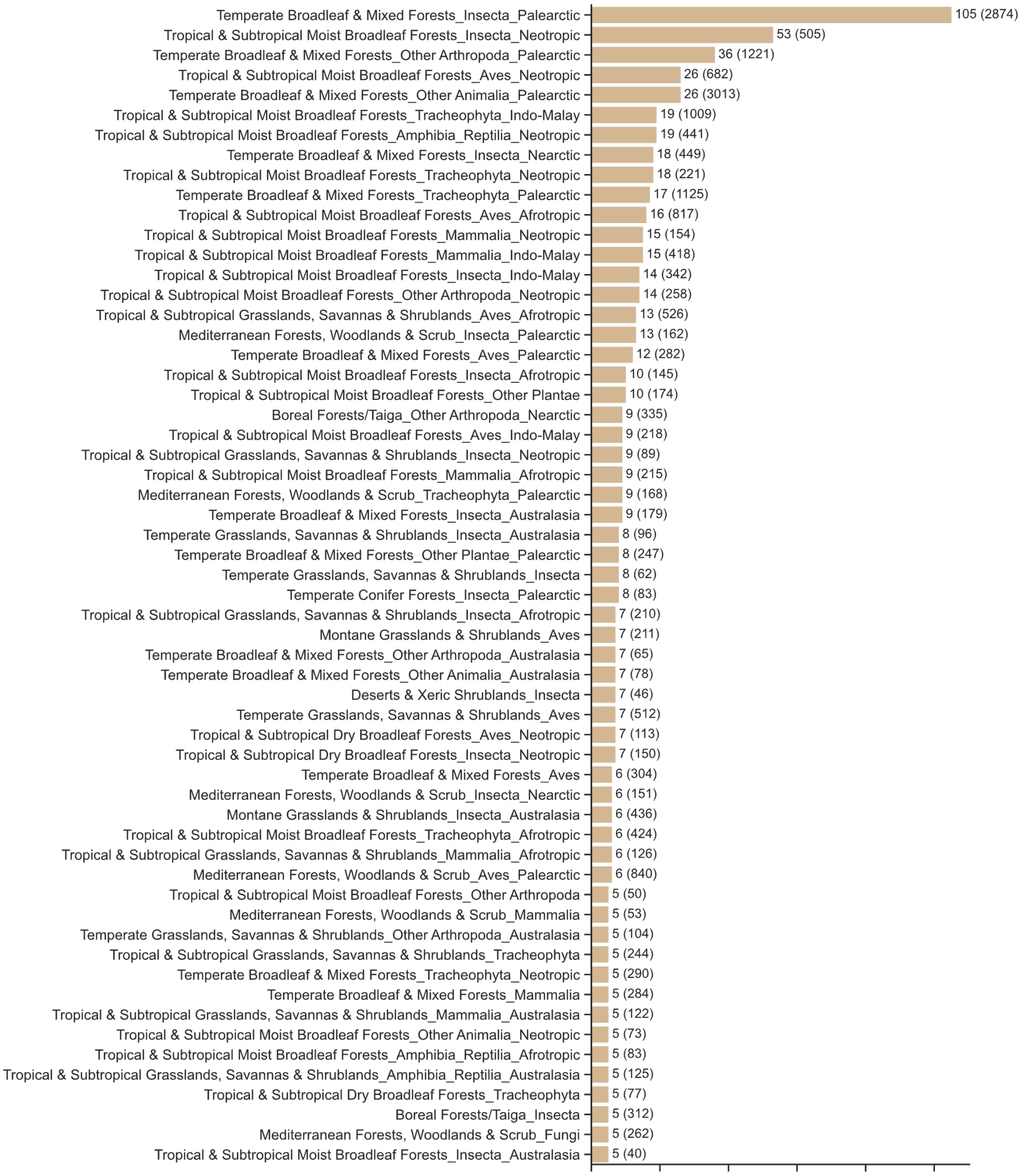
**Extended Data Fig. 2 | Country-level data coverage. a-b**, Number of source studies (**a**) and sampling sites (**b**) per country included in the training data. Countries with no data are shown in grey.

**Extended Data Table 1 | List of variables used across the different models**. These are divided into three categories: i) human impacts (used in all models), ii) environmental drivers (used in the BHMs), and iii) differences between site-pairs (used in the beta diversity models). The Haversine spatial distances are based on the sampling site coordinates from PREDICTS. The Gower environmental distance was calculated using the min and max temperature of the coldest and warmest month, the precipitation of the driest and wettest month, and elevation, all at the 1 km$^2$ scale. Repeated information from the row above is indicated by -.

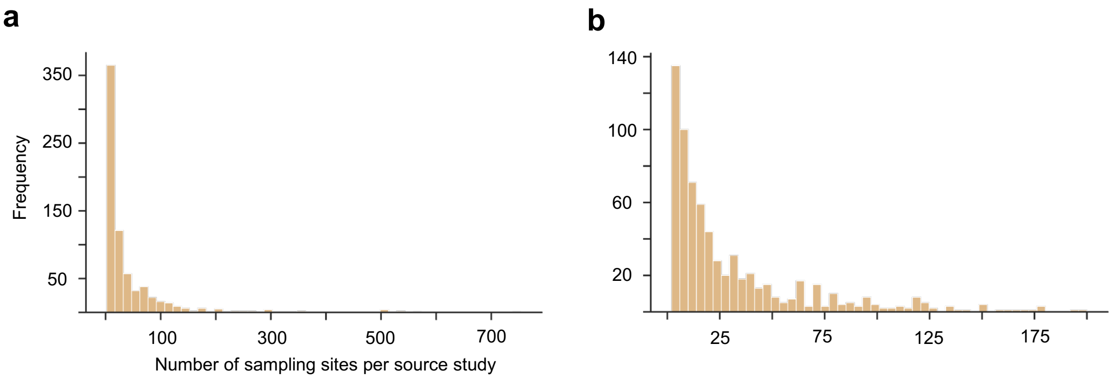
| Model variables | Spatial scale | Source | Spatial resolution | Temporal resolution |
|---|---|---|---|---|
| *i) Human impacts (all models)* | | | | |
| Primary vegetation, light use | Sampling site | PREDICTS | Varying | Sampling date |
| Primary vegetation, intense use | - | - | - | - |
| Secondary vegetation, minimal use | - | - | - | - |
| Secondary vegetation, light use | - | - | - | - |
| Secondary vegetation, intense use | - | - | - | - |
| Cropland, minimal use | - | - | - | - |
| Cropland, light use | - | - | - | - |
| Cropland, intense use | - | - | - | - |
| Pasture, minimal use | - | - | - | - |
| Pasture, light use | - | - | - | - |
| Pasture, intense use | - | - | - | - |
| Urban, all use intensities | - | - | - | - |
| Mean population density (log) | 1, 50 km$^2$ | GPW | 1 km$^2$ | Yearly |
| Mean road network density (log) | - | gROADS | Varying | Static |
| *ii) Environmental (BHMs)* | | | | |
| Annual mean temperature | 1 km$^2$ | BioClim | 1 km$^2$ | 1970–2000 avg |
| Temperature seasonality | - | - | - | - |
| Annual precipitation | - | - | - | - |
| Precipitation seasonality | - | - | - | - |
| Elevation | 1, 10 km$^2$ | EarthEnv | 1 km$^2$ | Static |
| Terrain roughness index | - | - | - | - |
| *iii) Site-site differences (beta models)* | | | | |
| Difference in (log) population density | 1 km$^2$ | GPW | 1 km$^2$ | Yearly |
| Difference in (log) road density | - | gROADS | Varying | Static |
| Haversine spatial distance | n/a | *Calculated* | n/a | n/a |
| Gower environmental distance | - | - | - | - |



**Extended Data Fig. 3 | Response variable distributions. a-b**, Data distribution of the response variables for alpha diversity, measured as the geometric mean abundance per site (**a**), and beta diversity, measured as the Bray-Curtis similarity between pairs of sites and reference sites within a study (**b**). Note that the distributions are slightly different in the BHMs, since some studies sampled multiple species groups, which leads to additional observations in the data.



**Extended Data Fig. 4 | Hierarchical groups in the BTMs.** Number of source studies and sampling sites (in parenthesis) in each final biogeographic-taxonomic group included in the BTMs. Some groups are at the biome-taxonomy-realm level, while others have been rolled up to the biome-taxonomy level. Groups that were rolled up to the overall population mean level have been excluded.



**Extended Data Fig. 5 | Sampling sites per study.** **a**, Distribution over the number of sampling sites per source study in the analysis. **b**, The same distribution but capped at 200 sites, to better show the distribution of the majority of source studies. In the data, 18.8% of studies have 5 sites or less, while 60.9% have 25 or less.