Actionable inference for biodiversity change hinges on representative data and model design

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. 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 Alliance, 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 authors:
- Jakob Nyström (jakob.nystrom@ebc.uu.se)
- Tobias Andermann (tobias.andermann@ebc.uu.se)

Abstract

Amidst the global 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 performance in spatial projections has not been systematically tested. W evaluate this using PREDICTS data, finding that, despite land-use impacts in line with previous research, there is a challenging gap between effect size inference and prediction. We find that mixed models with study attributes as random effects – common in biodiversity meta-analysis and indicators – exhibit low predictive accuracy, driven by reliance on highly averaged fixed effects. Ecologically structured models that replace study random effects with biome, realm, and taxonomic parameters, show improved but still modest results in sampled contexts. Yet, performance when extending predictions to other contexts remains low, due to distribution shifts in environmental factors and conditional responses. Our results highlight tensions between high-resolution predictor availability and the granularity at which responses can be reliably estimated. While models are essential for informed conservation efforts, their applicability is fundamentally constrained by data availability. Countries with extensive data can build higher-fidelity national indicators, but accelerated and systematically structured data collection is needed to support data-poor regions with localized, actionable biodiversity 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 health and prosperity depends[5]. The recently updated monitoring framework[6] of the Global Biodiversity Framework[7] (the GBF-MF) underscores the crucial role of biodiversity indicators to provide insights for nature monitoring, conservation and restoration[8]. In parallel, companies are accelerating efforts to understand nature-related dependencies, impacts, and risks[9–12]. Demand is therefore growing for robust, globally consistent indicators for assessing biodiversity change and its drivers[13].

Although biodiversity data repositories accumulate increasing amounts of species observations, lingering geographic and taxonomic gaps and biases make comprehensive assessment challenging[14–16]. Statistical models that estimate and predict how biodiversity relates to human pressures and environmental factors have the potential to fill such gaps, for use in model-based monitoring[17]. Global indicators like the Biodiversity Intactness Index (BII)[18–20], Mean Species Abundance (MSA) from the GLOBIO model[21,22], and the Biodiversity Habitat Index (BHI)[23,24], have been widely adopted for biodiversity assessment, monitoring, and scenario analysis, in the public and private sectors[1,2,6,13,25]. These indicators are built around the concept of biodiversity intactness[26], whereby the biodiversity of an area is estimated relative to comparable, ecologically intact reference sites. Outputs are spatially explicit, global maps of normalized intactness levels at high spatial resolutions (300 m to 1 km)[20,22,23]. They constitute a key use case for global biodiversity models, and their adoption underscores the value of modeling to support policy and decision-making[8,13,17,27].

Clearly, such large-scale modeling comes with challenges. In the absence of global networks for biodiversity data collection, models have to ingest information from many heterogeneous source studies, collated into meta-databases[28–30]. These data contain valuable ecological signal and structure, but also idiosyncrasies related to individual study scope, design, and sampling. Further, they tend to reflect the aforementioned gaps and biases in data availability. This prompts questions about model generality for use in prediction: the extent to which inferences from a sample apply to the sampled ecological context at large, and when extrapolating to other contexts[31]. To our knowledge, global pressure-response models have not been systematically evaluated for out-of-sample predictive performance[20,32]. Validation has focused on model fit[20,33] and parameter robustness across data subsets[18,34], which is important but not sufficient for assessing generality. At a local scale, researchers have found that not only can global biodiversity models perform poorly[33,35], but so do local models when extending predictions to other ecological conditions[36–38]. A global model scope, with clear data limitations, makes evaluation more challenging[20] but equally important[32].

One important question is how model structure and ecological granularity influences generality, in face of heterogeneous and sparse data. Mixed effects models[39] have routinely been applied to meta-databases for effect size estimation[31,40] and spatial projections[18,22]. For example, the BII quantifies fixed effects of land-use, human population density, and roads, averaged across all geographies and taxa[19,20], where random effects account for variation between studies. In the BHI, retention of biodiversity under land-use change is estimated in a similar way[23,24]. GLOBIO fits separate mixed models for land-use and several other pressures, for vertebrates and plants[22]. Such meta-analytic models have generated important insights about human impacts for conservation policy. Yet, their applicability for large-scale spatial predictions remains to be

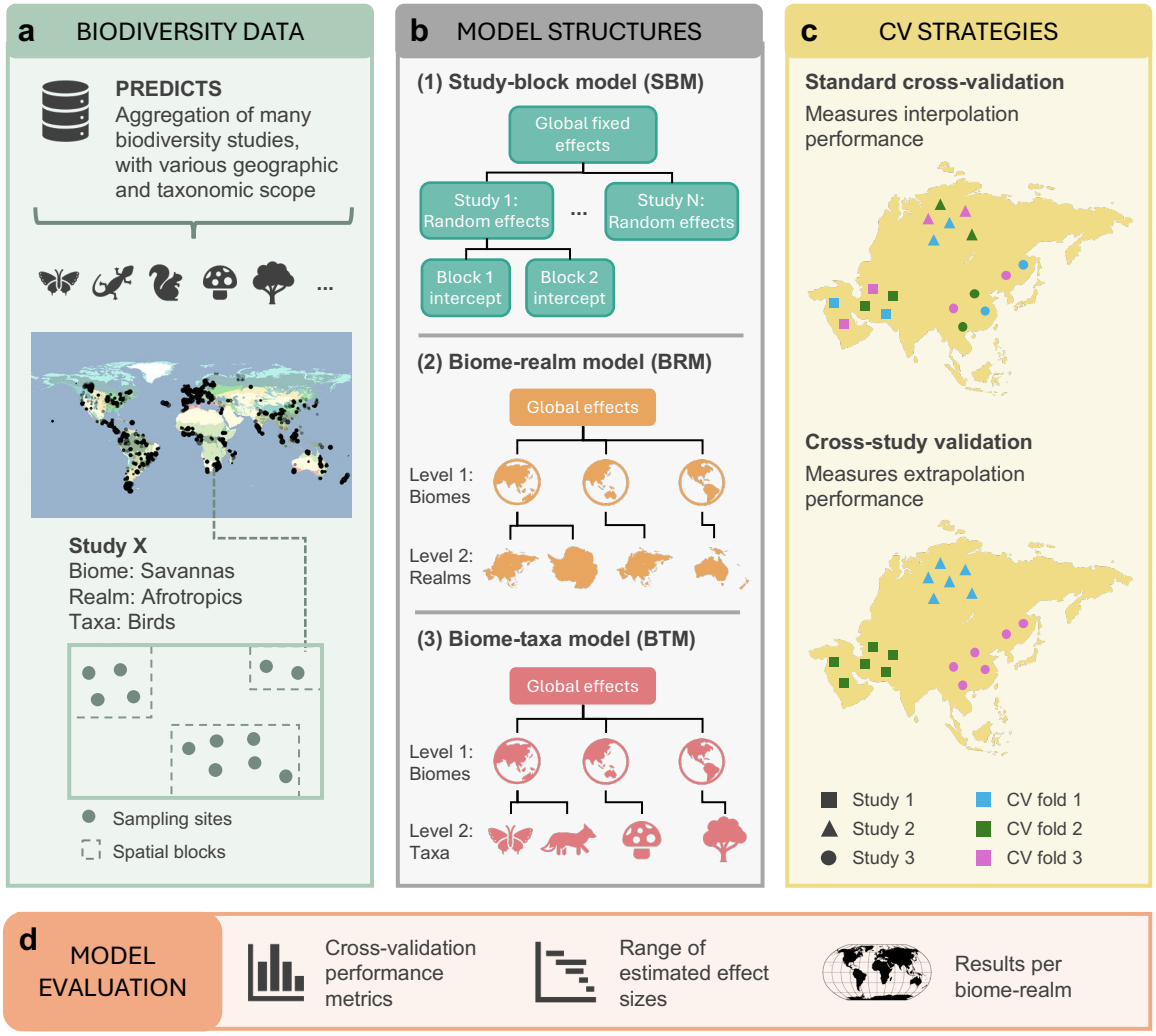


Fig. 1 | Main study components. **a**, Biodiversity and land-use data from the PREDICTS database were joined with data on human population density, road network density, bioclimatic factors, and topography. Geometric mean abundance was used to quantify site-level alpha diversity, while beta diversity was calculated as the Bray-Curtis similarity between ecologically intact reference sites and other sites. **b**, We trained three statistical pressure-response models: 1) A study-block model (SBM) with human pressures as fixed effects, where study-level random effects accounted for individual study variation. 2) An ecologically structured, hierarchical model with varying parameters per biome-realm (biome-realm model, BRM). 3) A structurally identical model where biomes instead were combined with taxonomic groups (biome-taxa model, BTM). **c**, Two cross-validation strategies were used: For interpolation, folds were generated by splitting data at the site level ('standard' CV). For testing extrapolation, folds were split such that all sites from a study only appeared in a single fold (cross-study validation). **d**, Models were assessed using rank correlation (Spearman ρ), residual-based R^2, and mean absolute error (MAE).

tested. For relevance in decision-making[19], granular model outputs are desirable since many ecological processes operate on local scales[19,24,41]. However, although environmental and human impact data are globally available at high spatial resolutions, it remains an open question how well biodiversity responses can be predicted at relevant scales using global models. One limitation of study random effects is that they cannot be used for out-of-sample predictions, potentially limiting model generality, if the fixed effects only explain a small portion of the

variation in the data[20]. Several studies have found notable differences in biodiversity responses across biogeographic regions[34,40], taxonomic groups[35,40], and more local scales[33]. These findings suggest that explicitly accounting for biogeographic and taxonomic context in global models could improve generality, but it is not clear to what extent this holds.

In this study, we investigate the generality of different model structures for estimating biodiversity intactness (key study components shown in Fig. 1). We leverage global data from the widely-used PREDICTS database[28], which has comparatively broad taxonomic coverage and serves as the basis of several indicators[18,22,23] (see Extended Data Fig. 1 and Extended Data Fig. 2 for model data coverage). Generality is operationalized as model accuracy when making spatially explicit, out-of-sample predictions in sampled ecological contexts (hereafter referred to as interpolation) as well as in other contexts (extrapolation). The model structures are implemented to deal with the heterogenous data differently, each with advantages and drawbacks. The first model uses study-level random effects and fixed effects that are averaged across all data. In the second model we substitute the study-based structure for a hierarchy that combines biomes and realms. The rationale is to learn differentiated ecological parameters that can be used for out-of-sample predictions, unlike study-specific effects. In the third model, biomes are combined hierarchically with a broad taxonomic grouping of species. By contrasting these models, we can explore implications of model design on the accuracy of global biodiversity predictions, in light of the current data landscape.

Results

We used geometric mean abundance (GMA)[42] to quantify site-level alpha diversity, normalized to a 0–1 scale across studies[20]. After filtering out studies with less than ten sites, the alpha diversity dataset consisted of 24,861 sites from 445 studies (see Extended Data Fig. 1 and Extended Data Fig. 3a,c). For beta diversity, we calculated the Bray-Curtis (BC) compositional similarity[42,43] between ecologically intact reference sites – consisting of minimally used primary vegetation – and other sites within a given study[20]. Due to the large number of possible site pairs, a balanced subsample was used, after removing studies with fewer than three reference sites. In total, this gave us 35,196 site-pairs from 184 studies (Extended Data Fig. 2 and Extended Data Fig. 3b,d)). Model covariates capturing human impacts included land-use[28], population density[44], and road network density[45] (see Extended Data Table 1). We also included covariates for temperature and precipitation[46], and elevation and terrain roughness[47], to capture broad environmental gradients beyond land-use and land-cover. The beta diversity models additionally included pairwise differences in human population and road density[20]. Here, environmental gradients were encoded as the Gower multivariate distance between sites. The models were implemented in a Bayesian hierarchical framework, to leverage statistical strength across groups and handle overparameterization in cases of sparse data[48–50].

A summary of model performance is shown in Fig. 2 (with fold-wise results in Extended Data Table 2 and calibration plots in Extended Data Fig. 6). In terms of alpha diversity (Fig. 2a), the model with random effects for studies and spatial blocks within studies (SBM) had a mean Spearman rank correlation of just 0.14 and near-zero R^2, when evaluated using standard cross-validation (CV). The biome-realm model (BRM) had substantially better rank correlation (0.42) and R^2 (0.16), while MAE showed only a small difference. Here, the hierarchical structure contained biomes at the first level, subdivided by biogeographic realms at the second level (see Extended Data Fig. 4). In the biome-taxa model (BTM), the second hierarchical level used a

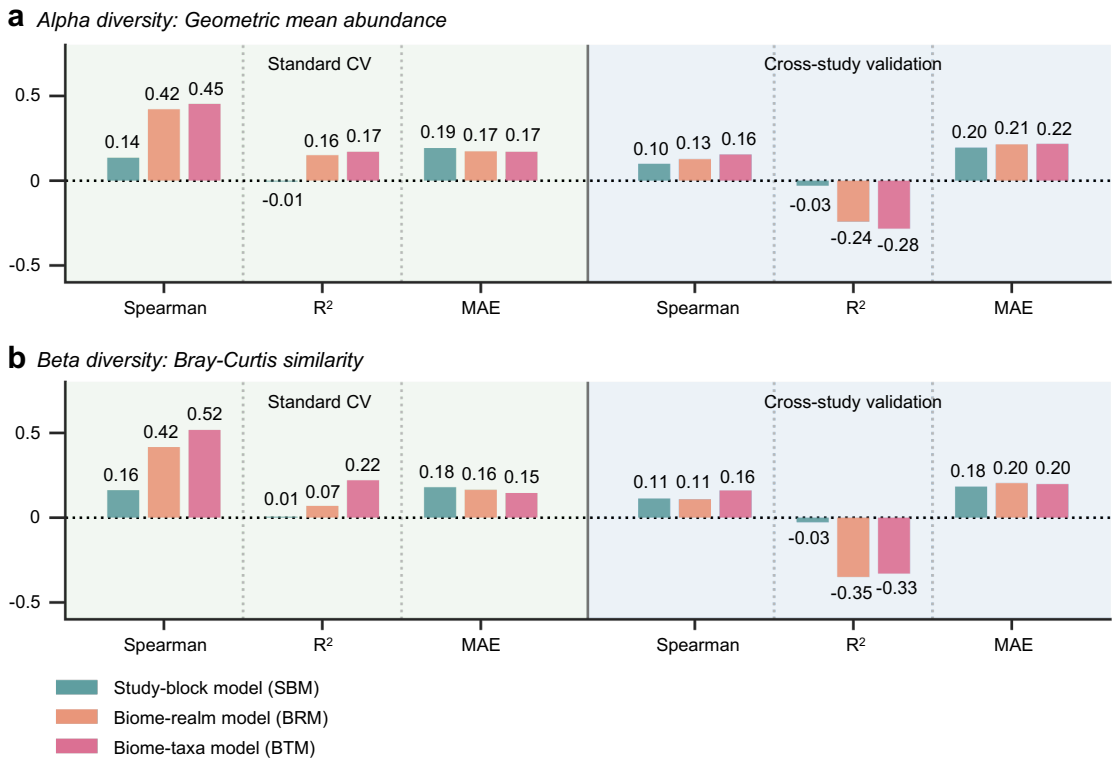


Fig. 2 | Model performance summary. Predictive performance quantified using Spearman rank correlation, residual-based R^2, and mean absolute error (MAE) on all test points. **a**, Alpha diversity: Normalized site-level geometric mean abundance. **b**, Beta diversity: Bray–Curtis similarity between pairs of ecologically intact reference sites and other sites, within each study.For Spearman and R^2, higher is better; for MAE, lower is better. Green bars show the study-block model (SBM), orange bars the biome-realm model (BRM), and red bars the biome-taxa model (BTM). Light green backgrounds show results for standard CV, while light blue represents cross-study validation.

taxonomic grouping of plants, vertebrates, invertebrates, fungi and other taxa (see Extended Data Fig. 5). The corresponding rank correlation and R^2 were somewhat higher, at 0.45 and 0.17. In standard CV, folds were generated using stratified random sampling of sites across all studies, to assess interpolation accuracy within sampled contexts (light green panels). In the BRM and BTM, to reduce the risk of optimistic interpolation results, only groups with at least five studies and 100 sampling sites, plus five reference sites for beta diversity models, utilized group-level parameters for predictions. Smaller groups were instead 'rolled up' to the level above. That said, the BRM and BTM assume some degree of data representativeness across the studies of each ecological group in the final prediction hierarchy.

Turning to cross-study validation[51] (Fig. 2, light blue panels), which evaluates extrapolation to other contexts by fully separating underlying studies between folds, results were markedly worse. The BRM rank correlation was only 0.13 while R^2 was negative at -0.24. Similarly, the BTM metrics dropped to 0.16 (Spearman) and -0.28 (R^2). The SBM performance did not deteriorate as much, relative to its weaker starting point. For all models, the increases in MAE were small compared to the decreases in the other metrics. For beta diversity (Fig. 2b) the relative patterns were similar. The BRM and BTM did substantially better in terms of interpolation, but again saw a large drop in performance metrics during cross-study validation. The cross-study validation

procedure seemed to provide effective guardrails against the overall performance evaluation being too optimistic. We also tested the inclusion of study-level intercepts and pressure slopes as controls, to avoid ecological parameters absorbing any major study-specific effects (shown in Extended Data Table 2). The results were inconclusive; in standard CV, rank correlation and R^2 were worse for both diversity metrics and models, with the only improvement seen in cross-study validation of the beta diversity model. This was likely due to weak parameter identifiability from high model complexity in relation to sparse study- and group-level data.

It should be noted that the structural differences between the models have some implications on the intercomparison. In the SBM, the fixed effects were estimated based on data from all sites, covering the full spatial and taxonomic scope of the data. The site-level predictions, generated by applying those parameters to local human pressures, were then evaluated against site-level observed values that represent a very small share of that full ecological scope. The BRM and BTM were trained on data where the response variables were split by biomes, realms and broad taxonomic groups, making evaluation of predictions more specific to each ecological context. While this is a non-negligible difference, the approach used here was the most appropriate way of evaluating the models when used in predictive tasks.

Drivers of relative model performance

The better interpolation performance of the BRM and BTM, compared to the SBM, was largely explained by the decomposition of variance explained[52] (Fig. 3a,b). The SBM suffered from low attribution of variance to observable fixed effects, relative to the study-level random effects (similar to previous studies[20]). Since cross-validation simulates model performance on unseen sites, the random effects cannot be used when making out-of-sample predictions. In contrast, the BRM and BTM learned differentiated response parameters for different ecological groups. Since these were based on general biogeographic and taxonomic classifications, they could also be used to predict out-of-sample observations. This dichotomy is further illustrated by Fig. 3c-e, which shows the model parameters used for predictions, in relation to underlying cross-study heterogeneity, for the alpha diversity models (corresponding results for beta diversity in Extended Data Fig. 7). Through their flexibility, the ecologically structured models were able to capture a substantial portion of this variation (Fig. 3d-e), compared to the SBM fixed effects (Fig. 3c). Based on previous studies[33–35,40] it can expected that differentiated, group-level driver responses improve predictions within sampled contexts. Still, their true generality of course depends on the representativeness of the underlying data. To avoid extreme group-level parameters, the BRM and BTM were regularized through weakly informative priors[49], while the study-block model (SBM) priors were kept loose to not constrain study heterogeneity.

For some covariates, the range of group-level parameters extended beyond the study-level ranges (Fig. 3d-e). This can happen if there is limited data support to reliably estimate group-level effects, making the model susceptible to fitting noisy observations. Despite differences, all models exhibited strong regression towards the mean (Extended Data Fig. 6), producing overbiased predictions for small observed values, and vice versa. The SBM predictions, in particular, were more or less flat across all observed values (Extended Data Fig. 6a). In line with previous studies[3,4,18,40], most land-use types had negative average estimated effects on biodiversity compared to minimally used primary vegetation (Fig. 3c). The impacts of human population and road density were more ambiguous, possibly because these variables are temporally static. Additionally, we note that the BRM and BTM global posterior means often had smaller magnitudes compared to the SBM fixed effects (Fig. 3c-e). It could imply that the models

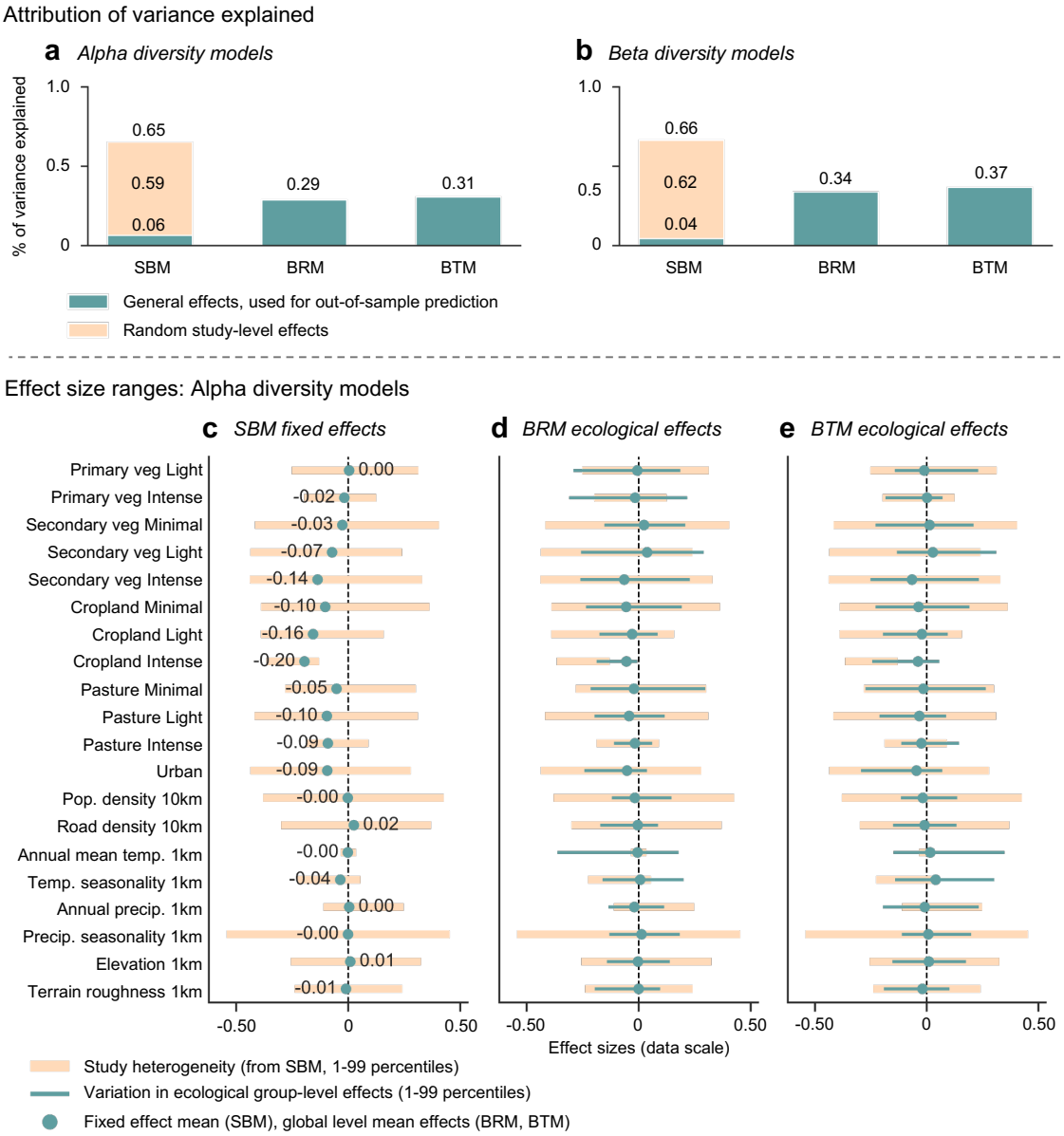


Fig. 3 | Variance explained and effect size ranges. a,**b**, Variance explained (variance-based R^2) of the training data, for the alpha diversity (**a**) and beta diversity (**b**) models. Green bars represent non-study effects that can be used for out-of-sample predictions, while yellow bars show the contribution of study-specific random effects. The BRM and BTM have no yellow bars since they include no study effects in the main results. **c**-**e**, Estimated ranges of model parameters in the alpha diversity models. Yellow bars indicate study heterogeneity, the spread of study-level random effects. These are based on the SBM, and shown in all panels for comparability. Effect sizes are expressed on the scale of the observed data (0–1). **c**, Green circles denote SBM fixed effects, which are averages across all studies. **d**,**e**, Green bars show the range of group-level parameters of the BRM (**d**) and BTM (**e**), respectively. Mean values are left out to not clutter the plots.

attributed relatively little variation to the global level, compared to the lower hierarchical levels. The large drops in BRM and BTM extrapolation accuracy compared to interpolation (Fig. 2) were caused by distribution shifts between training and test data, forcing the models to make out-of-distribution predictions[37,38]. In Fig. 4a,b we calculated the covariate Gower distance[53]

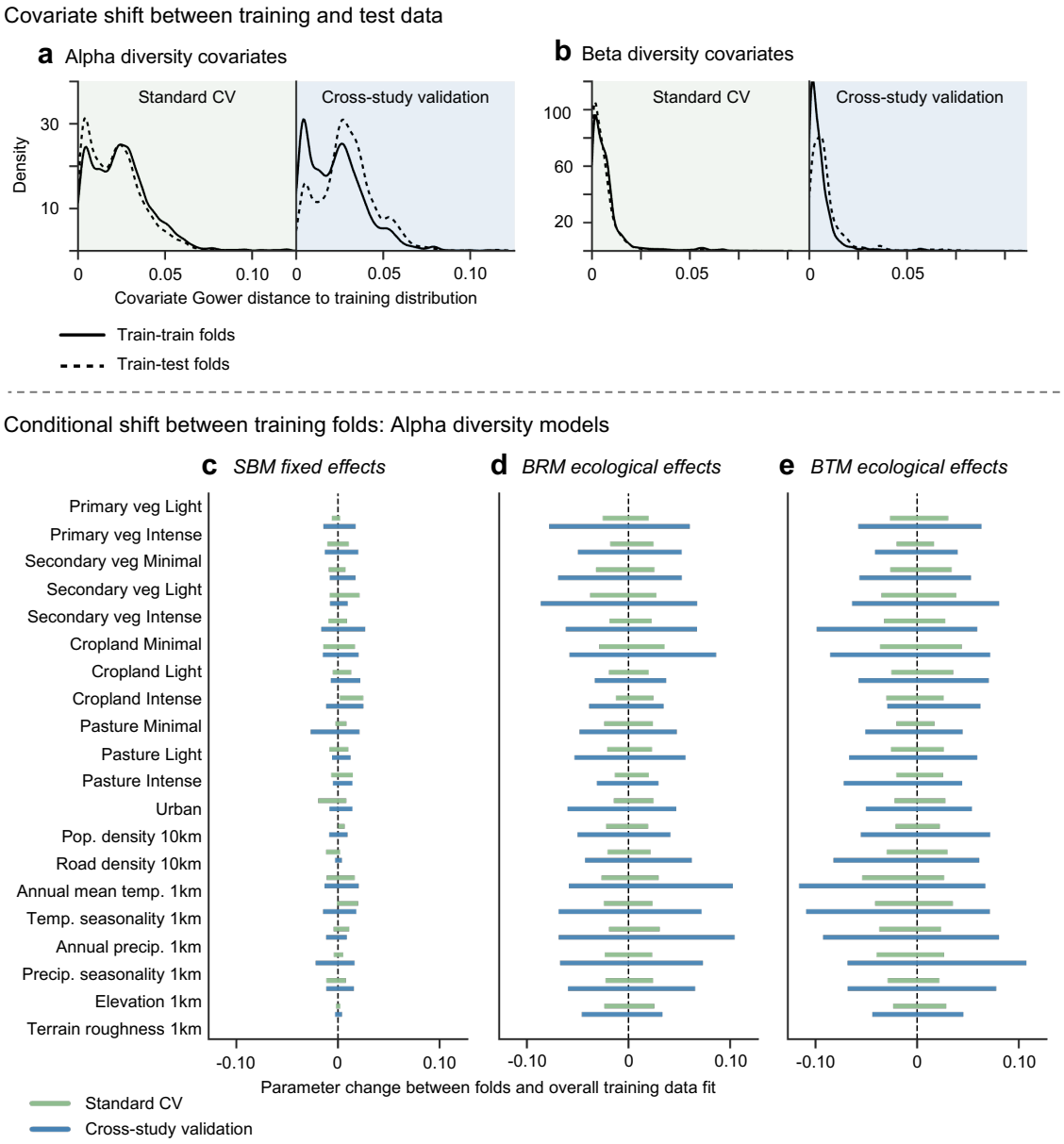


Fig. 4 | Covariate and conditional shifts. **a**,**b**, Covariate shifts between training and test distributions for the alpha diversity (**a**) and beta diversity (**b**) datasets. These were obtained by calculating the median Gower distance of model covariates between (i) a sample of 2,000 points from the whole dataset, and (ii) up to 200 points from the training data (solid line) and test data (dotted line) of each biome-realm and fold. Light green panels show standard CV, while light blue represent cross-study validation. **c**-**e**, Change in estimated model parameters between each training fold fit, and the fitted parameters for the whole dataset, for the alpha diversity SBM (**c**), BRM (**d**) and BTM (**e**).

between the fold-wise training and test distributions, and the overall data distribution, per biome-realm. For alpha diversity (Fig. 4a) and standard CV, the distribution of test data closely followed the training data, relative to the overall data (light green panel). Under cross-study validation, the test data clearly shifted in covariate space, due to changing environmental conditions (light blue panel). Since environmental differences were encoded as a single variable in the beta diversity models, calculated within individual studies, the shift was less pronounced (Fig. 4b).

Shifts in conditional responses seemed to be the strongest driver of low extrapolation performance, especially for the BRM and BTM (Fig. 4d,e). The fitted parameters of each training fold varied substantially more during cross-study validation (blue bars) compared to standard CV (green bars). These shifts occurred because the cross-study validation induced environmental and taxonomic separation of training and test data, by keeping entire studies contained to single folds (beta diversity results shown in Extended Data Fig. 8). The flexibility of the BRM and BTM, an advantage for interpolation, here led to overfitting to studies seen in training. This is clearly illustrated by the model calibration plots in Extended Data Fig. 6b (showing the BRM predictions). In contrast, the rigidity of the SBM meant that parameters did not change much between folds compared to fitting on the overall data (Extended Data Fig. 6a). Consequently, predictions were more consistent between interpolation and extrapolation, on an overall low level.

Results for biome-realms

In Fig. 5a,b, we show heatmaps of relative performance (using the Spearman rank correlation) across the regional biomes that constitute the hierarchical groups in the BRM (based on the standard CV scenario). Since data is limited in many groups, the results are indicative, but still highlight some clear performance differences. On group-level, interpolation performance appears overall higher for alpha diversity (Fig. 5a) compared to beta diversity (Fig. 5b), in contrast to the mean results across all the data (Fig. 2). One reason is that the alpha diversity dataset contains more than twice as many studies than the beta diversity dataset, resulting in better coverage per biome-realm (Extended Data Fig. 1 and Extended Data Fig. 2).

There were noticable geographic differences in performance, but it is hard to draw conclusions at this level. To investigate this, we regressed the group-level rank correlations on key data attributes of each group (Fig. 5c,d). For alpha diversity, we found that data extent (total number of sites) and taxonomic depth (orders per study and species per study and order) were positively associated with model rank correlation, although number of sites was not significant at the 95% level. On the contrary, data heterogeneity, here represented by the number of studies and range of environmental conditions (Gower covariate distances) had negative associations with group-level performance. The beta diversity results (Fig. 5f) showed some interesting differences, with number of studies and sites per study contributing positively to interpolation performance. That the BC similarity data were generated from pairwise reference site comparisons, within the context of individual studies, likely explains the reversed effect. Environmental heterogeneity (Gower distance) did not have a negative effect here, likely for the same reason, and also that it is included as a model variable. In general, these findings are indicative and warrant further analysis beyond the scope of this study.

Discussion

In this study, we evaluated the predictive performance of biodiversity intactness models using a global dataset, finding clear limits to spatially explicit model predictions (Fig. 2). While some generalization in terms of interpolation performance could be attained within well-sampled ecological contexts, none of the models achieved transferability, measured as extrapolation accuracy, across contexts. Given the demand for global biodiversity indicators[13], this systematic assessment fills an important knowledge gap. The results highlight a tension between the valuable high-level insights that can be drawn from global models and their challenges in accurately predicting local relationships, considering the high spatial resolution of many biodiversity data products. To some extent, higher predictive accuracy can be reached through targeted choices in

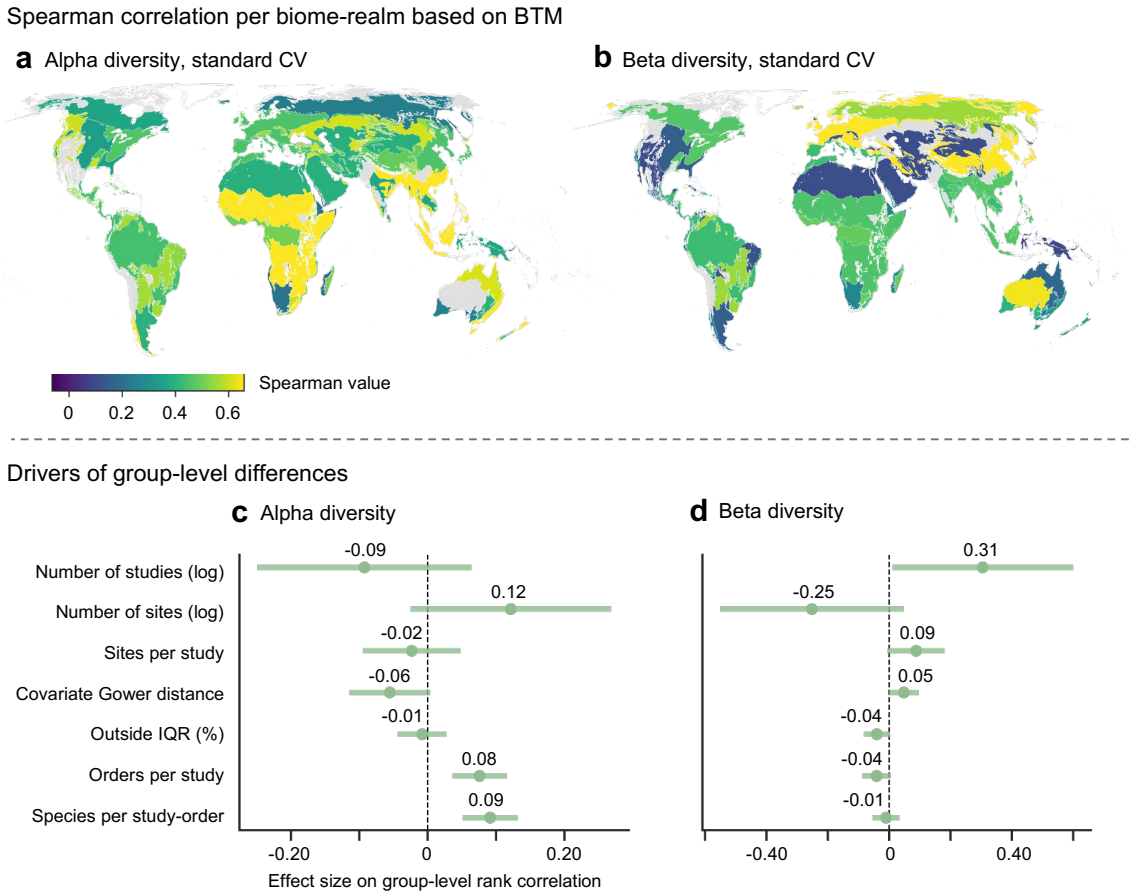


Fig. 5 | Results across biome-realms. **a**,**b**, Global heatmaps of Spearman rank correlation in each biome-realm, based on predictions from the BRM. Groups with less than ten observations have been excluded. Group-level results have been winsorized at the 5th and 95th percentiles to prevent outliers from skewing the heatmap gradient. **c**,**d**, Estimated effects of regressing group-wise rank correlations on potential drivers of relative performance, for alpha diversity (**c**) and beta diversity (**d**). Circles show mean coefficients and bars represent 95% confidence intervals. The Gower distance is the median distance of all observations to their ten nearest neighbors within each biome-realm. Observations outside the IQR refer to values of the response variable.

model design. Yet, there are substantial limits to prediction at global scales, partially resulting from the heterogenous and incomplete nature of currently available biodiversity data. While sobering, these insights can be used to further develop biodiversity models and their critical role in policy and decision-making.

In terms of interpolation, the models with explicit ecological structures (BRM and BTM) did better than the pure random effects model (SBM), in particular on ranking of higher and lower diversity sites (Fig. 2). This is logical since these models attribute more data variation to effects that can be used for out-of-sample predictions (Fig. 3). Still, interpolation performance, based on rank correlation and R^2, was overall low for all models. This emphasizes the challenge of predictive modeling across large biogeographic and taxonomic scales. Further, none of the models did well when extrapolating to new contexts. The BTM, which included both biogeographic and taxonomic structure, did slightly better than the SBM and BRM in terms of rank correlation (Fig. 2). However, the BRM and BTM both made large prediction errors, resulting in negative

R^2 values. The SBM showed much smaller differences between standard CV and cross-study validation, due to its rigid fixed effects structure. Model flexibility is clearly advantageous for interpolation, but does not necessarily translate to better extrapolation. Low transferability driven by covariate and conditional distribution shifts (Fig. 4) is a known issue in biodiversity modeling[36–38]. Our results highlight the magnitude of this issue for models trained on global, heterogenous data when applied to local contexts. The differences across across biome-realms (Fig. 5a,b) suggest that the quantity and heterogeneity of underlying model data plays an important role for relative model performance (Fig. 5c,d). This implies that lingering data gaps and biases on a global level[14,15] put fundamental constraints on large-scale biodiversity models. The results align with previous evaluations of global models in more specific local contexts[33,35].

Scientifically robust indicators are essential for achieving the goals of the GBF[6,8]. Our results corroborate that land-use change is a strong driver of biodiversity loss (Fig. 3c), in line with previous analyses[3,4,18,20,40]. However, low model accuracies emphasize the challenging gap between effect size inference and spatially explicit predictions, where explanatory power is not necessarily an indicator of predictive power[54,55]. The granularity at which conditional responses can be estimated is equally important as the availability of high-resolution environmental and pressure data. Regional biomes have been proposed as a suitable level for model-based monitoring[40], but the limited BRM predictive performance challenge that assumption. The same holds for using broad taxonomic groups like plants, vertebrates, and invertebrates[18,34,40]. To effectively support decision-making, model limitations like the ones shown in this study should be fully transparent to users, with explicit performance evaluation being crucial[32]. Data products where spatial resolution is not matched by predictive accuracy can incur risks of sub-optimal decision-making. The results of this study suggest that global, taxonomically broad models, trained on currently available data, are most appropriately used for directional, high-level insights on biodiversity drivers and patterns. While that might already be clear to some researchers and users, it is seemingly at odds with indicator outputs at fine spatial scales[18,22,23]. This can potentially lead to overconfidence in inferred patterns and trends.

Countries with high-quality national biodiversity data can overcome several limitations by implementing national-scale models and indices[6,8], using the established methodology of established indicators. Although this could reduce inter-country comparability, it would greatly improve relevance for decision-making at appropriate scales. However, that still leaves data-poor regions with a lack of accurate biodiversity insights[56]. Existing data, such as structured GBIF sampling event datasets, can to some extent be used to expand meta-databases. That requires significant standardization efforts, though, and will not resolve overall biases in available data. The global community must therefore allocate resources to scale up collection of biodiversity data through standardized monitoring programs[14,57], leveraging cost-effective technologies like environmental DNA, bioacoustics and camera traps[58]. These sampling efforts should be guided by expected increases in model performance and reductions in uncertainty. Similarly, companies that want to ensure high-quality reporting and decision-making will often need to collect proprietary data to complement existing tools[12].

Some limitations of our study also point to important research priorities. Fairly assessing models with such a large scope is complex[20]; we already highlighted some reservations earlier. Combining standard CV with cross-study validation provides an indication of performance in different scenarios, but in reality the boundaries between interpolation and extrapolation are fuzzy. Cross-validation is the established method of evaluating predictive models, but it is hard to account for

every possible source of underlying data bias. Ecological groups with limited sampling could suffer from inflated interpolation performance, despite the use of rolled-up parameters, especially since the BRM and BTM does not contain explicit study-level control variables. There is a need for refined model evaluation frameworks that can handle large scales in multiple dimensions. This is crucial in order to determine whether a model is appropriate to apply for a particular use case, in a specific ecological context (spatially and taxonomically), and with a certain output resolution. This involves estimating data representativeness, model risk, and uncertainty, while accounting for possible distribution shifts between data used for training and the intended scope of application. In addition, some notion of what constitutes 'good enough' performance is required. This could, for example, be based on model skill compared to a naive baseline of mean or random predictions. Still, the exact metrics and thresholds to apply ultimately come down to informed scientific and user judgment, before applying a model or indicator.

Methods

Data sources

Biodiversity data: All biodiversity data – species presence or absence, and abundance – were obtained from the Projecting Responses of Ecological Diversity in Changing Terrestrial Systems (PREDICTS) project[28], a meta-database compiled from independent source studies and inventories. We combined the two publicly available releases of the database (from 2016 and 2022)[59,60] into one dataset. Before filtering (see further down), the combined dataset contained 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 PREDICTS data 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, some tropical areas in Latin America also have high coverage, while there are significant gaps in Africa and parts of Asia. See Extended Data Fig. 1 and Extended Data Fig. 2 for an overview of the PREDICTS data used in the modeling. In 255 of the source studies (pre-filtering), spatially adjacent sampling sites had been grouped into spatial blocks, which was utilized for the study-block model (SBM).

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[28]. 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[44], 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 population data represents extrapolated numbers for the years 2000, 2005, 2010, 2015, and 2020. Data on road networks were taken from the Global Roads Open Access Data Set, v1 (gROADS)[45], a global layer of joined country road networks adjusted for topology. 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[46], frequently used in species distribution models and other biodiversity studies. The data represent average values over the period 1970–2000. Data on elevation, slope and terrain roughness came from the EarthEnv repository[47], constructed from global digital elevation models derived from satellite imagery and LiDAR measurements.

The spatial resolution of the PREDICTS land-use data depends on the specific sampling extent of a given site in each source study[28] (which is only available for 13.1% of sampling sites). The resolution of the population density, bioclimatic and topographic variables is 30 arc-seconds (approximately 1 km^2 at the equator)[44,46,47]. The spatial accuracy of the road network data varies by country[45]. Manual classification of land-use type and intensity could have introduced inconsistencies within and between studies[28]. Since predominant land-use is a categorical attribute, there might be underlying differences in habitat conditions within the same class that are not observed.

Biodiversity metrics

We modeled two biodiversity metrics, the geometric mean abundance (GMA)[42], an alpha diversity metric, and the Bray-Curtis (BC) similarity[43], a beta diversity metric. Combined, they provide insights into the relative level of community richness, abundance, evenness, and compositional similarity between areas, and are suitable for detecting changes in biodiversity[42]. 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 site abundance, a more even distribution of abundance across several species 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 any of the 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 is highly unlikely). Since the focus of this study is on models used to estimate biodiversity intactness, the BC similarity is calculated between reference sites, consisting of minimally used primary vegetation sites, and all other sites within each study. This follows the approach used in the BII[20] and is similar to the implementation of MSA in GLOBIO[22].

One challenge with biodiversity 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 we found, following the BII[20], was to normalize the GMA values within each study to a common 0–1 scale, by dividing them by the respective within-study maxima:

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

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 $y_i > 1.5\,\text{IQR}$, within each study.

The distributions of the alpha diversity (GMA) and beta diversity (BC similarity) metrics are shown in Extended Data Fig. 3a-d. They are non-symmetric and quite heavily skewed to the

right. The shapes are similar, but the GMA distributions have greater means than the BC indices, since a compositional similarity metric to a greater extent reflects natural species 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. In addition, studies that only surveyed a small number of species are more likely to have zeros at the site level. However, there is no feasible way to construct a separate detection probability model[48,61] for such a heterogeneous dataset using the data that we have available. The inflation of zeros in the BC distribution is the result of 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 above, since every study will contribute a one (its maximum abundance site) to the overall data pool.

In addition to the requirement on abundance data, studies with fewer than ten sites were filtered out from the GMA dataset. Further, we required at least three minimally used primary vegetation reference sites in each study included in the BC dataset. Still, it should be noted that many source studies still contain very few sites (see Extended Data Fig. 3e,f). 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 ideally 25–50 sites per study would be required to produce somewhat continuous distributions of data at the study-level. Some of this is alleviated by pooling data across studies, but some biases can still persist.

The biome-taxa model (BTM) used taxonomic groups as part of the hierarchical model structure (see Extended Data Fig. 5). Since some larger studies had sampled species across several such taxonomic groups, sampling sites were consequently split into multiple observations. This only had a very minor effect on the number of observations and the distribution of response variables (see Extended Data Fig. 3a-d). The BTM alpha model had 25,104 observations in total, in comparison to the 24,861 sampling sites in scope, which also equaled the number of observations in the study-block (SBM) and biome-realm (BRM) models. Since the BC similarity dataset contained more than 400,000 unique site pairs after standard filtering, we used sub-sampling to obtain a tractable but representative dataset for model training and evaluation. The sampling was designed to retain all studies with at least three reference sites, while preventing large studies, with many sites and reference sites, from becoming too dominant. For each study s, we calculated the number of potential pairs $n_{sites} \times n_{ref}$ and a sublinear weight $w_s = \sqrt{n_{sites} \times n_{ref}}$, with w_{tot} denoting the sum of all study weights. An initial number of pairs was allocated to each study based on its actual number of pairs n_{pairs} and an overall target fraction $f = 0.15$ of the whole dataset. This was weighted by w_s and subject to lower and upper bounds k_{min}, k_{max}, such that $k_s = \max(k_{min}, \min(n_{pairs} \times f \times w_s/w_{tot}), k_{max})$. The threshold values were set to $k_{min} = 300$ and $k_{min} = 3,000$, respectively, to increase the contribution of smaller studies and limit the dominance of the largest ones. To balance reference sites and other sites within each study sample, we finally chose $\sqrt{k_{study}}$ non-reference sites and $k_{study}/\sqrt{k_{study}}$ reference site, to get approximately k_{study} sites in total. This resulted in a total of 35,196 site pairs that were consistent across all models.

Model covariates

Variables based on land-use[28], population density[44], and road network density[45] data, formed the core of all models in the study[20]. For land-use, the land-use class and its intensity was combined into a new set of categorical variables; records where this information was unknown were filtered out. 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, following previous work[20]. 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 10 km^2 scales, log transformed to reduce skewness, were also used as human pressure variables. In addition, the following bioclimatic and topographic variables were included: annual mean precipitation and temperature, temperature and precipitation seasonality, elevation, and terrain roughness, all at 1 km^2 resolution. A complete list of model covariates can be found in Extended Data Table 1. We did not include any interaction effects due to the small number of observations in many studies (and to some extent, hierarchical groups), implying that the models were already highly parameterized, and even overparameterized to some extent.

To derive the continuous variables, the coordinates of each sampling site were first projected from global EPSG:4326 to local UTM format, before generating equal area sampling windows corresponding to the model resolution for that variable, and then reprojecting to global format in order to extract data from the global raster layers. The equal-area projections ensured that the calculated values were 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). 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[62]. The population densities were matched to the sampling year of each site; the other covariates were static layers.

For the beta diversity (BC) models, 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 multivariate environmental (Gower) distance[53] was included to control for natural compositional turnover[20]. The Gower environmental distance was calculated using the annual mean temperature, temperature seasonality, annual precipitation, precipitation seasonality, min and max temperature of the coldest and warmest month, the precipitation of the driest and wettest month, elevation, and terrain roughness, all at the 1 km^2 scale. All continuous variables were standardized prior to model training and evaluated for collinearity. The highest pair-wise correlations were 0.61 (terrain roughness and precipitation seasonality) and 0.58 (annual mean temperature and temperature seasonality).

Model structures and implementation

Likelihood function: The GMA and BC distributions are non-symmetric and bounded between 0 and 1, with substantial inflation of zeros and some inflation of ones in the GMA case (Extended Data Fig. 3a-d). This suggests that a zero-inflated or zero-one-inflated beta distribution would be the most appropriate likelihood function to describe the data[63]. The beta distribution is 0–1 bounded and can have a more or less symmetrical shape depending on its parameters. 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 (due to the inter-study normalization) that did not warrant explicit modeling either. We therefore assumed a regular beta likelihood function, with a logit link function, for all models[64].

Modeling framework: Although previous global biodiversity intactness modeling efforts have utilized frequentist mixed effects models[18,20,22,23], we decided to implement the models in this study using a Bayesian hierarchical framework. This was a requirement for fitting the ecologically structured models (BRM and BTM), since the framework can effectively leverage statistical strength across groups and handle overparameterization in cases of sparse data[48–50], with regularization through weakly informative priors[49]. The models were implemented in `PyMC`[65], using the No-U-Turn Sampler (NUTS)[66] with a `numpyro`[67] backend. For all experiments, we ran four parallel sampling chains for stability, using 1,000 tuning samples that were discarded and 1,000 posterior draws. A high target acceptance rate of 0.95 was used to avoid problems when sampling from the complex posterior distribution. Due to the high number of small studies and ecological groups, it was hard to avoid some parameters with small effective sample sizes and $\widehat{R} > 1.01$[68]. While this could indicate issues in the sampling, the overall performance results were robust also when using fewer iterations than in the final runs.

Study-block model: The SBM model structure consisted of fixed effects, parameters representing average values over all studies in the dataset, and random effects, which quantify group-level variation around the fixed effects[39,69]. The study-level random 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 were estimated as an average of all geographies and taxa in the data[18–20].

We let $\boldsymbol{\beta}$ denote the fixed effects and let $\boldsymbol{\gamma}_s$ denote the study-level random effects, for the study s that site i belongs to. The spatial block intercepts are denoted by $\gamma_{sb(s)}$ for each block sb within study s. These different effects constitute a hierarchical structure, from the population of all studies to individual source studies, and from studies to blocks. As noted above, we used a beta likelihood with a logit link function $g(\cdot)$ across all models[64]. The mean of y_i, conditional on the covariates, is denoted by μ_i. The site-level regression model for GMA can then be written as

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{x}_i^\top \boldsymbol{\gamma}_s + \gamma_{sb(s)} = \eta_i,$$

where $\mu_i = g^{-1}(\eta_i) = e^{\eta_i}/(1 + e^{\eta_i})$. We fitted a full set of random slopes at the study level, such that the random effect covariates are the same as the fixed effect ones $\mathbf{x}_i$. The $\boldsymbol{\gamma}_s$ parameters were assumed to be uncorrelated and vary around the fixed effects with mean zero.

In the beta diversity model, y_{ij} denotes the BC similarity between some site i and a reference site j, within the context of a study s. In addition to the regular 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 these difference terms, we can write the beta diversity model as

$$g(\mu_{ij}) = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\beta}^\Delta + \mathbf{x}_i^\top \boldsymbol{\gamma}_s + \mathbf{z}_{ij}^\top \boldsymbol{\gamma}_s^\Delta + \gamma_{sb(s)}.$$

For the dispersion parameter of the beta distribution likelihood, we sampled a raw scale parameter $\sigma_{\text{raw}} \sim \text{Beta}(a, b)$, using $a = 2, b = 12$ for the GMA model and $a = 2, b = 20$ for the BC similarity model. We then defined a mean-dependent dispersion $\sigma(\mu_{i,t}) = \sigma_{\text{raw}} \sqrt{\mu_{i,t}\,(1 - \mu_{i,t})}$ and used it to parameterize the beta likelihood:

$$y_{i,t} \sim \text{Beta}(\mu_{i,t}, \sigma(\mu_{i,t}))\,.$$

The priors of the SBM were kept lose in order to not constrain the estimated study heterogeneity, but constrained enough for the sampling to run efficiently.

Biome-realm and biome-taxa models: The SBM model structure is centered around the data collection process, with studies that sampled sites which were sometimes organized into spatial blocks. In the BRM and BTM, we instead designed the hierarchical structure based on the biomes and realms where each site is located, and the taxonomic groups that were sampled at that site. In both models, this was implemented as a two-level hierarchy. Biomes constituted the first hierarchical level in both models. These are distinct in their environmental conditions and underlying biodiversity patterns, which suggests that the effect sizes of both natural and anthropogenic drivers will be different between them[34,40]. In the BRM, the second level subdivided biomes by biogeographic realms, to form regional biomes[40] (see Extended Data Fig. 4). In the BTM, the second level instead consisted of broad taxonomic groups, which previous studies have also suggested differ in their responses[22,35,40]: plants, vertebrates, invertebrates, fungi and other taxa (see Extended Data Fig. 5).

We let $b = 1, \ldots, B$ denote biomes, $r = 1, \ldots, R$ biogeographic realms, and $t = 1, \ldots, T$ taxonomic groups. Further, $r(b)$ indicates the partition of a given biome into realms, and $t(b)$ the equivalent split of biome data by taxonomic groups. Note that realms can span multiple biomes and vice versa, so this is not a unique one-to-one mapping (and the same of course holds for the taxonomic groups). For the group-level model parameters, intercepts are denoted by α and other parameters by $\boldsymbol{\beta}$. We provide details on the BRM model formulation below; the BTM model followed the same structure and is therefore omitted for brevity. The alpha diversity (GMA) model for i, in the regional biome $r(b)$ can be written as

$$g(\mu_i) = \alpha_{r(b)} + \mathbf{x}_i^\top \boldsymbol{\beta}_{r(b)}.$$

Like the SBM, we assumed normal and uncorrelated priors for the model parameters, with half-normal priors on the variance terms. At the population-level, the following hyperpriors were used:

$$\mu_\alpha \sim \mathcal{N}(0.35, 0.25^2),\ \mu_\beta^{(k)} \sim \mathcal{N}(0, 0.18^2) \text{ for } k = 1, \ldots, K,$$

where the K is the number of regression parameters. These priors define distributions over the population-level model parameters, which sit at the top of the hierarchy. The priors at the first hierarchical level (biomes) were drawn from the hyperpriors in the following way:

$$\begin{aligned} \alpha : \sigma_\alpha &\sim \text{Half-Normal}(0.15^2),\ \alpha_b \sim \mathcal{N}(\mu_\alpha, \tau_b \cdot \sigma_\alpha^2) \\ \beta : \sigma_\beta &\sim \text{Half-Normal}(0.07^2),\ \beta_b^{(k)} \sim \mathcal{N}(\mu_\beta, \tau_b \cdot \sigma_\beta^2) \end{aligned}$$

The scaling factor, $\tau_b = \min(\ln(n_b) - 1, 0)$, adapted the amount of regularization – through the prior variance – on the group-level model parameters as a function of the number of studies in

each biome-taxa group. In a similar way, the priors at the second level (biome-realm) inherited from their respective parent groups at the first level, but with tighter priors on the slope variances:

$$\alpha : \sigma_\alpha \sim \text{Half-Normal}(0.15^2),\ \alpha_{r(b)} \sim \mathcal{N}(\alpha_b, \tau_{r(b)} \cdot \sigma_\alpha^2),$$
$$\beta : \sigma_\beta \sim \text{Half-Normal}(0.05^2),\ \beta_{r(b)}^{(k)} \sim \mathcal{N}(\beta_b^{(k)}, \tau_{r(b)} \cdot \sigma_\beta^2).$$

The numeric prior values above were chosen based iterative prior predictive checks, to obtain a prior predictive distribution reasonably similar to the observed data (Extended Data Fig. 3a,c) and achieve a relatively high degree of model regularization. Next, the beta diversity model can be expressed as

$$g(\mu_{ij}) = \alpha_{r(b)} + \mathbf{x}_i^\top \boldsymbol{\beta}_{r(b)} + \mathbf{z}_{ij}^\top \boldsymbol{\beta}_{r(b)}^\Delta.$$

The prior structure was identical to the alpha diversity model, but we assumed slightly different hyperpriors due to the lower mean and longer tail of the BC similarity distribution (Extended Data Fig. 3b,d):

$$\mu_\alpha \sim \mathcal{N}(0.25, 0.25^2),\ \sigma_\alpha \sim \text{Half-Normal}(0.18^2).$$

Since many biogeographic and taxonomic groups contained only a few studies and sites (Extended Data Fig. 3e,f), there was a risk that the estimated group-level parameters would become too specific to a few studies, rather than generally applicable to the group at large. To prevent this from producing overoptimistic accuracy numbers, especially for interpolation, we implemented a 'roll-up' scheme for out-of-sample predictions. Here, biome-realm or biome-taxa groups with less than five studies and 100 sampling sites in each training fold, plus five references sites for beta diversity, were assigned the parameters of their respective parent groups at the biome level for prediction. If the threshold was still not met at the biome-taxa level, the population-level posterior parameters were used.

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 model runs, with stratified sampling using biomes as strata. The first approach, using 'standard' CV, assessed how well models could make predictions in environmental and taxonomic contexts similar to the data that they were trained on. In other words, this evaluated model interpolation performance (generalizability) from sample to sampled population. 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 generally split among several folds, creating overlap in studies, but not sites, between folds. In the BTMs, fold assignment was still done at site level, keeping multiple taxonomic units for a given site in the same fold, if present. For the beta diversity runs, sampling into folds was based on the non-reference site in each pair, meaning the same reference site used as baseline could be part of both training and test data. We deemed this to be acceptable given the relatively few reference sites in many studies, which would otherwise have led to very unbalanced folds.

The second approach, based on cross-study validation[51], was used to assess how well the models could make predictions in new contexts, that is their extrapolation performance (transferability). 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 (biomes), in which case it could appear in more than one fold. The cross-study validation

approach led to a clear spatial and environmental separation of training and test data, while also reflecting taxonomic and methodological differences between source studies. Although the numbers of studies per fold were roughly equal, one consequence of the cross-study 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 used three complementary performance metrics to evaluate the models. The Spearman rank correlation provides a measure of how good a model is at ranking higher and lower diversity sites, which crucial for real-world decision-making.

$$\rho_s = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(\widehat{r}_i - \bar{\widehat{r}})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(\widehat{r}_i - \bar{\widehat{r}})^2}},$$

where $r_i = \text{rank}(y_i)$ and $\widehat{r}_i = \text{rank}(\widehat{y}_i)$. Residual-based R^2 evaluates the squared errors of the model predictions compared to a baseline of predicting the mean of the test data, where the squared terms penalize large deviations.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

The final performance metric was the mean absolute error (MAE). Since the response variables were on a 0–1 scale, this metric can be interpreted as the average percentage point deviation between predicted and observed values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right|.$$

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. Goodsell, R., Granqvist, E., Christiaen, C. & Ronquist, F. Local Data Matters: Improving Biodiversity Risk and Impact Assessment through a Data Quality Focus. Preprint at https://ecoevorxiv.org/repository/view/11213/ (2025).
13. Burgess, N. D. *et al.* Global Metrics for Terrestrial Biodiversity. *Annu. Rev. Environ. Resour.* **49,** 673–709 (2024).
14. Hughes, A. C. *et al.* Sampling Biases Shape Our View of the Natural World. *Ecography* **44,** 1259–1269 (2021).
15. Chapman, M. *et al.* Biodiversity Monitoring for a Just Planetary Future. *Science* **383,** 34–36 (2024).
16. Boyd, R. J., Powney, G. D. & Pescott, O. L. We Need to Talk about Nonprobability Samples. *Trends Ecol. Evol.* **38,** 521–531 (2023).
17. Purvis, A. Bending the Curve of Biodiversity Loss Requires a 'Satnav' for Nature. *Phil. Trans. R. Soc. B* **380,** 20230210 (2025).
18. Newbold, T. *et al.* Global Effects of Land Use on Local Terrestrial Biodiversity. *Nature* **520,** 45–50 (2015).
19. Newbold, T. *et al.* Has Land Use Pushed Terrestrial Biodiversity beyond the Planetary Boundary? A Global Assessment. *Science* **353,** 288–291 (2016).
20. 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).
21. Alkemade, R. *et al.* GLOBIO3: A Framework to Investigate Options for Reducing Global Terrestrial Biodiversity Loss. *Ecosystems* **12,** 374–390 (2009).
22. Schipper, A. M. *et al.* Projecting Terrestrial Biodiversity Intactness with GLOBIO 4. *Glob. Change Biol.* **26,** 760–771 (2020).

23. Harwood, T. *et al.* BHI v2: Biodiversity Habitat Index: 30s Global Time Series. CSIRO https://doi.org/10.25919/TT2T-H452 (2022).
24. Hoskins, A. J. *et al.* BILBI: Supporting Global Biodiversity Assessment through High-Resolution Macroecological Modelling. *Environmental Modelling & Software* **132,** 104806 (2020).
25. Damania, R. *et al. Nature's Frontiers: Achieving Sustainability, Efficiency, and Prosperity with Natural Capital* (World Bank, Washington, DC, 2023).
26. Scholes, R. J. & Biggs, R. A Biodiversity Intactness Index. *Nature* **434,** 45–49 (2005).
27. Zurell, D. *et al.* Predicting the Way Forward for the Global Biodiversity Framework. *Proc. Natl. Acad. Sci. USA* **122,** e2501695122 (2025).
28. 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).
29. Dornelas, M. *et al.* BioTIME: A Database of Biodiversity Time Series for the Anthropocene. *Global Ecology and Biogeography* **27,** 760–786 (2018).
30. Sabatini, F. M. *et al.* sPlotOpen – An Environmentally Balanced, Open-Access, Global Dataset of Vegetation Plots. *Global Ecology and Biogeography* **30,** 1740–1764 (2021).
31. Spake, R. *et al.* Improving Quantitative Synthesis to Achieve Generality in Ecology. *Nat. Ecol. Evol.* **6,** 1818–1828 (2022).
32. Martin, P. A., Green, R. E. & Balmford, A. The Biodiversity Intactness Index May Underestimate Losses. *Nat. Ecol. Evol.* **3,** 862–863 (2019).
33. Jung, M. *et al.* Local Factors Mediate the Response of Biodiversity to Land Use on Two African Mountains. *Animal Conservation* **20,** 370–381 (2017).
34. Phillips, H. R. P., Newbold, T. & Purvis, A. Land-Use Effects on Local Biodiversity in Tropical Forests Vary between Continents. *Biodivers Conserv* **26,** 2251–2270 (2017).
35. De Palma, A. *et al.* Predicting Bee Community Responses to Land-Use Changes: Effects of Geographic and Taxonomic Biases. *Sci Rep* **6,** 31153 (2016).
36. Roberts, D. R. *et al.* Cross-validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* **40,** 913–929 (2017).
37. Meyer, H. & Pebesma, E. Predicting into Unknown Space? Estimating the Area of Applicability of Spatial Prediction Models. *Methods. Ecol. Evol.* **12,** 1620–1633 (2021).
38. Meyer, H. & Pebesma, E. Machine Learning-Based Global Maps of Ecological Variables and the Challenge of Assessing Them. *Nat. Commun.* **13** (2022).
39. Harrison, X. A. *et al.* A Brief Introduction to Mixed Effects Modelling and Multi-Model Inference in Ecology. *PeerJ* **6,** e4794 (2018).
40. Bevan, P. A. *et al.* Regional Biomes Outperform Broader Spatial Units in Capturing Biodiversity Responses to Land-use Change. *Ecography* **2025,** e07318 (2024).
41. Maris, V. *et al.* Prediction in Ecology: Promises, Obstacles and Clarifications. *Oikos* **127,** 171–183 (2018).
42. Santini, L. *et al.* Assessing the Suitability of Diversity Metrics to Detect Biodiversity Change. *Biol. Conserv.* **213,** 341–350 (2017).
43. Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* **27,** 325–349 (1957).
44. 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).

45. 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).
46. 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).
47. Amatulli, G. *et al.* A Suite of Global, Cross-Scale Topographic Variables for Environmental and Biodiversity Modeling. *Sci. Data* **5,** 180040 (2018).
48. Dorazio, R. M. Bayesian Data Analysis in Population Ecology: Motivations, Methods, and Benefits. *Popul. Ecol.* **58,** 31–44 (2016).
49. Lemoine, N. P. Moving beyond Noninformative Priors: Why and How to Choose Weakly Informative Priors in Bayesian Analyses. *Oikos* **128,** 912–928 (2019).
50. Gelman, A. Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics* **48,** 432–435 (2006).
51. Bernau, C. *et al.* Cross-Study Validation for the Assessment of Prediction Algorithms. *Bioinformatics* **30,** i105–i112 (2014).
52. 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).
53. Gower, J. C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **27,** 857–871 (1971).
54. Shmueli, G. To Explain or to Predict? *Stat. Sci.* **25,** 289–310 (2010).
55. 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).
56. Clements, H. S. *et al.* A Place-Based Assessment of Biodiversity Intactness in Sub-Saharan Africa. *Nature,* 1–9 (2025).
57. Gonzalez, A. *et al.* A Global Biodiversity Observing System to Unite Monitoring and Guide Action. *Nat. Ecol. Evol.* **7,** 1947–1952 (2023).
58. Hardwick, B. *et al.* LIFEPLAN: A Worldwide Biodiversity Sampling Design. *PLoS ONE* **19,** e0313353 (2024).
59. Hudson, L. *et al.* The 2016 Release of the PREDICTS Database V1.1. Natural History Museum https://doi.org/10.5519/J4SH7E0W (2023).
60. Contu, S. *et al.* Release of Data Added to the PREDICTS Database. Natural History Museum https://doi.org/10.5519/JG7I52DG (2022).
61. 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).
62. Goldewijk, K. K. Three Centuries of Global Population Growth: A Spatial Referenced Population (Density) Database for 1700–2000. *Popul. Environ.* **26,** 343–367 (2005).
63. Ospina, R. & Ferrari, S. L. P. Inflated Beta Distributions. *Stat Papers* **51,** 111–126 (2010).
64. Ferrari, S. & Cribari-Neto, F. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics* **31,** 799–815 (2004).
65. Abril-Pla, O. *et al.* PyMC: A Modern, and Comprehensive Probabilistic Programming Framework in Python. *PeerJ Comput. Sci.* **9,** e1516 (2023).
66. Hoffman, M. D. & Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15,** 1593–1623 (2014).
67. Phan, D., Pradhan, N. & Jankowiak, M. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. Preprint at https://arxiv.org/abs/1912.11554 (2019).

68. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.-C. Rank-Normalization, Folding, and Localization: An Improved R-Hat for Assessing Convergence of MCMC. Preprint at https://arxiv.org/abs/1903.08008 (2019).
69. Pinheiro, J. C. & Bates, D. M. *Mixed-Effects Models in S and S-PLUS* Ch. 1 (Springer, 2000).

Acknowledgments

We thank Alice Hughes for feedback on conceptual framing and preliminary results. We also thank Robert Goodsell, Emma Granqvist, Adrian Baggström, Mahwash Jamy, Vun Wen Jie and one anonymous reviewer for reading and providing feedback on the manuscript.

This research was supported by the SciLifeLab & Wallenberg Data Driven Life Science Program (grant: KAW 2020.0239) and the Swedish Research Council (2023-05366).

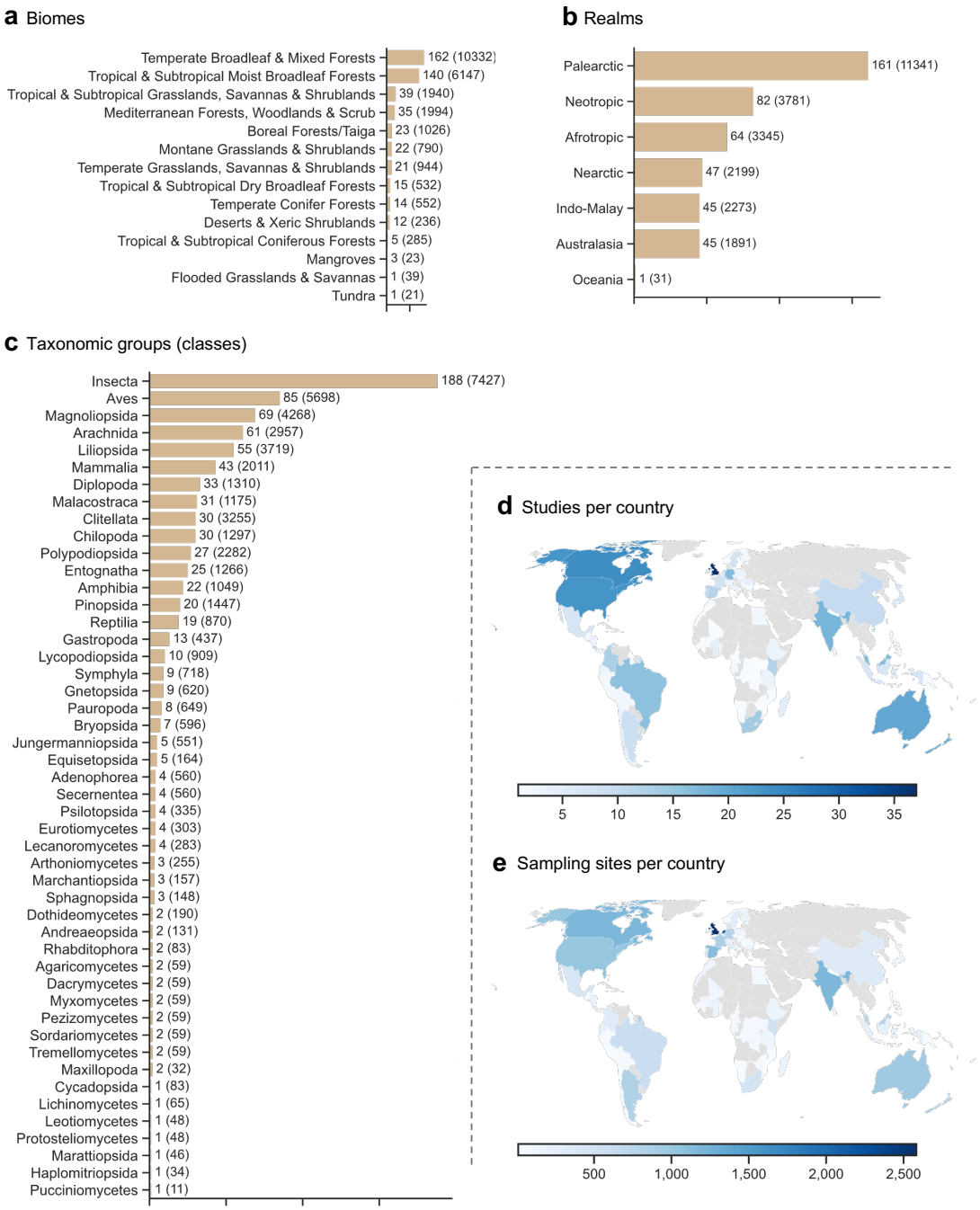
Author contributions

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.

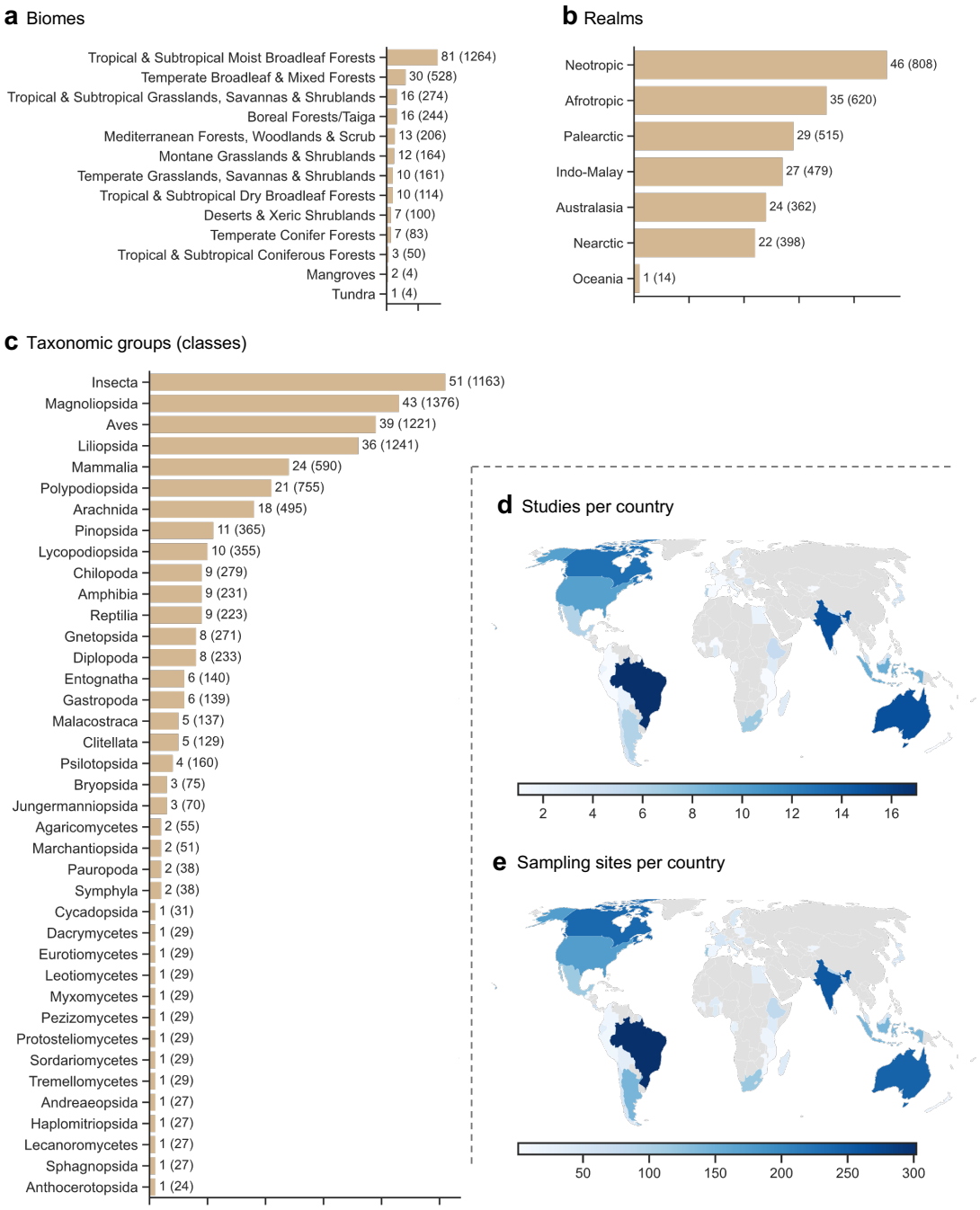
Code availability

The code for this study can be found at: https://github.com/j-nystrom/biodiversity-impacts.

Extended Data



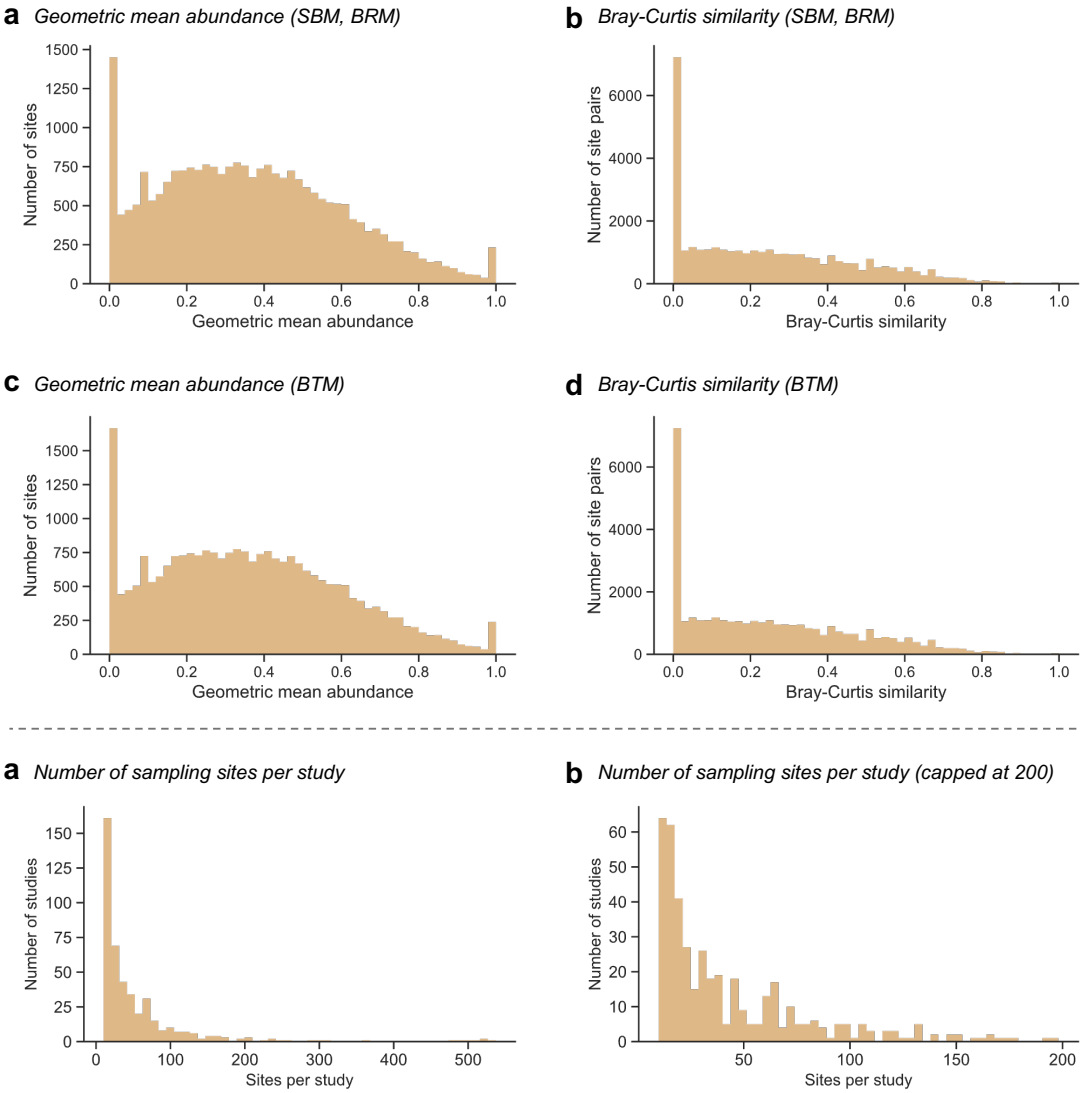
Extended Data Fig. 1 | Data coverage, alpha diversity models. a-c, Number of source studies and sampling sites (in parenthesis) in the alpha diversity model data, split by biomes (**a**), realms (**b**), and taxonomic classes (**c**). **a,b**, Number of source studies (**a**) and sampling sites (**b**) per country in the data. Countries with no data are shown in grey. This is after filtering out studies with less than ten sampling sites.



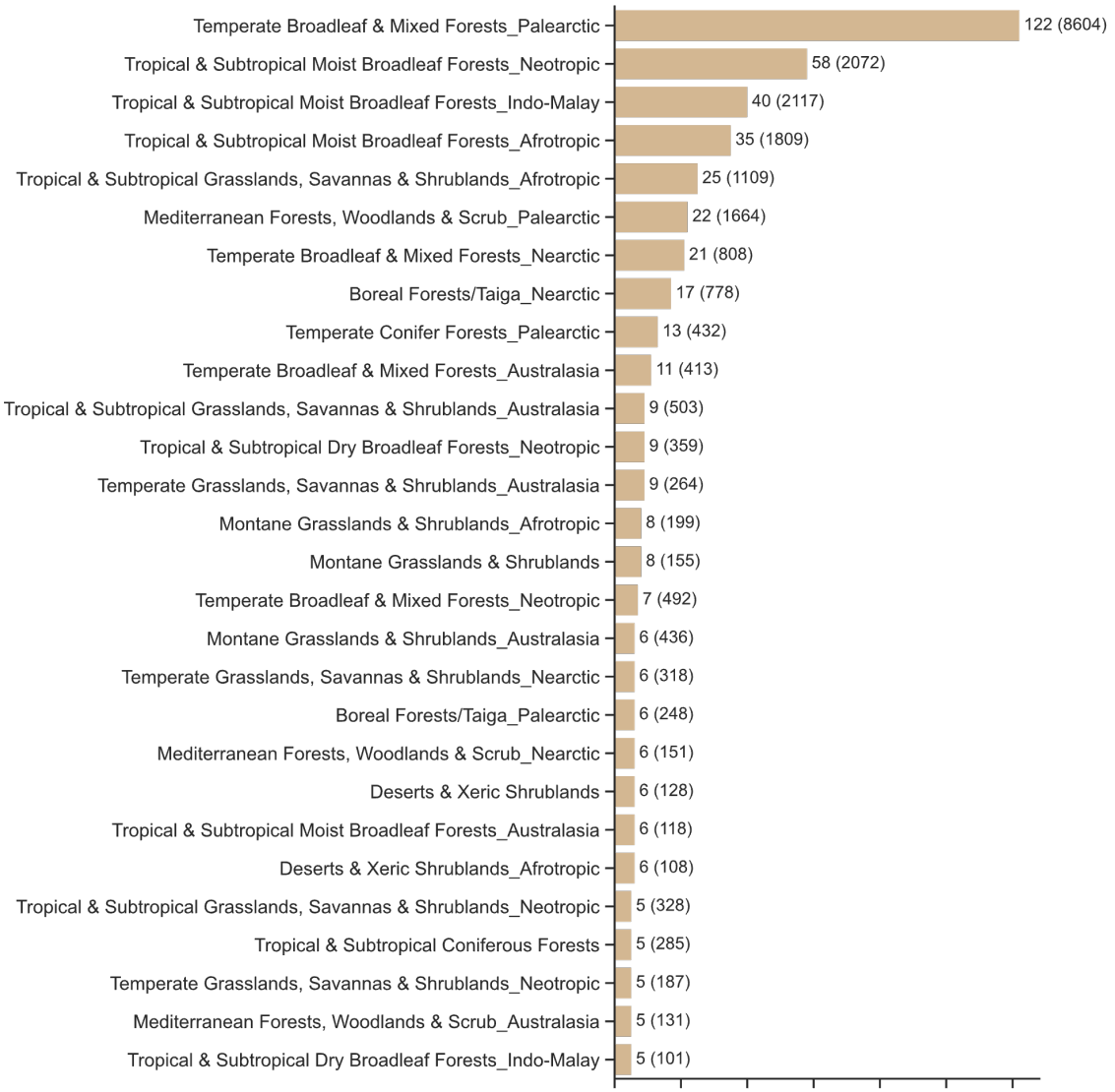
Extended Data Fig. 2 | Data coverage, beta diversity models. a-c, Number of source studies and sampling sites (in parenthesis) in the beta diversity model data, split by biomes (**a**), realms (**b**), and taxonomic classes (**c**). **a,b**, Number of source studies (**a**) and sampling sites (**b**) per country in the data. Countries with no data are shown in grey. This is after filtering out studies with less than ten sampling sites and three minimally used primary vegetation reference sites.

Extended Data Table 1 | List of variables used across the different models. These are divided into three categories: i) human pressure variables, ii) environmental drivers, and iii) differences between site-pairs (used in the beta diversity models). The Gower environmental distance was calculated using the annual mean temperature, temperature seasonality, annual precipitation, precipitation seasonality, min and max temperature of the coldest and warmest month, the precipitation of the driest and wettest month, elevation, and terrain roughness, all at the 1 km^2 scale. Repeated information from the row above is indicated by -.

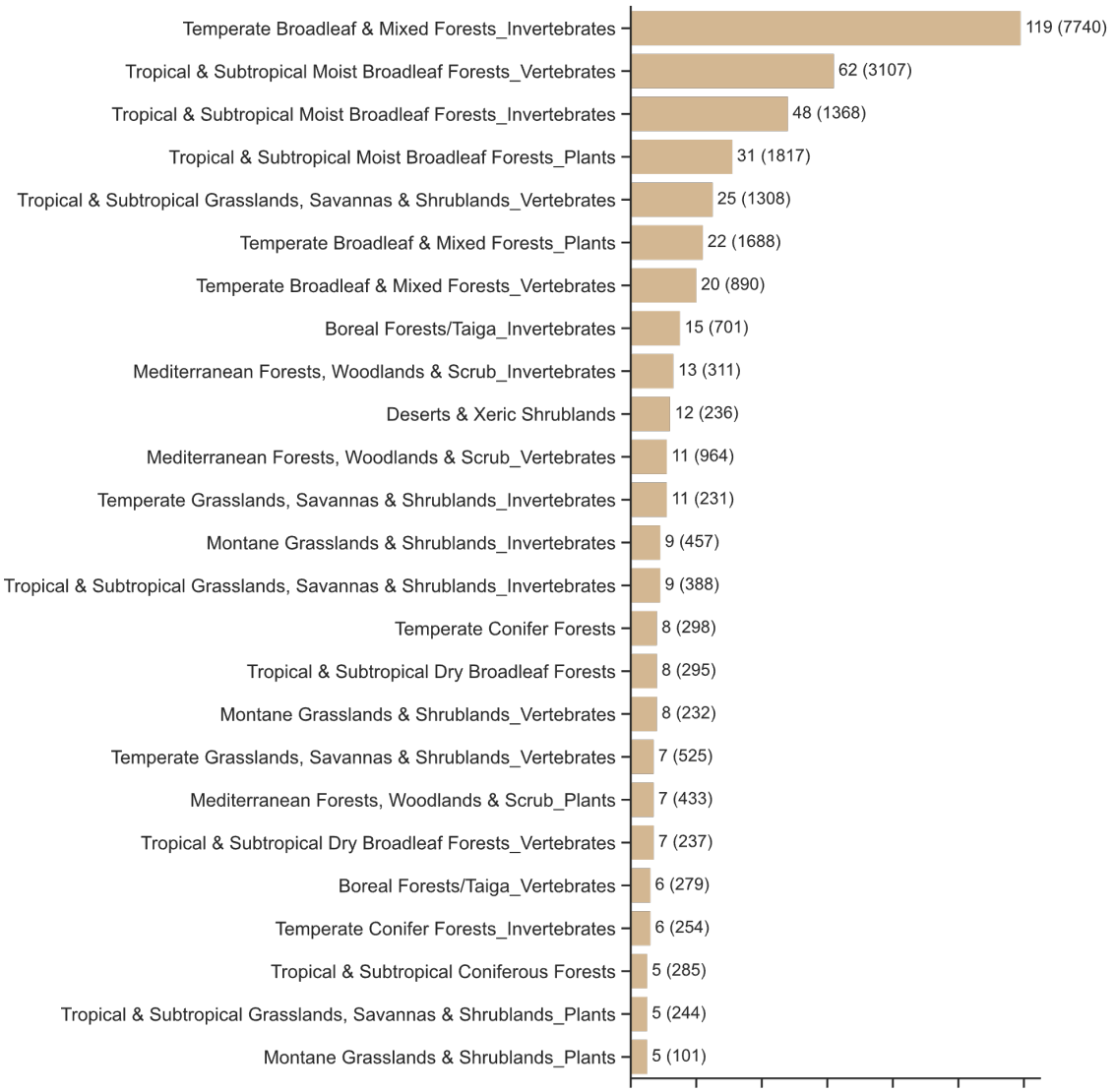
Model variables	Spatial scale	Source	Spatial resolution	Temporal resolution
i) Human pressures				
Primary vegetation, light use	Sampling site	PREDICTS	Varying	Sampling date
Primary vegetation, intense use	-	-	-	-
Secondary vegetation, minimal use	-	-	-	-
Secondary vegetation, light use	-	-	-	-
Secondary 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)	10 km^2	GPW	1 km^2	Yearly
Mean road network density (log)	-	gROADS	Varying	Static
ii) Environmental drivers				
Annual mean temperature	1 km^2	BioClim	1 km^2	1970–2000 avg
Temperature seasonality	-	-	-	-
Annual precipitation	-	-	-	-
Precipitation seasonality	-	-	-	-
Elevation	1 km^2	EarthEnv	1 km^2	Static
Terrain roughness index	-	-	-	-
iii) Site-site differences				
Difference in (log) population density	10 km^2	GPW	1 km^2	Yearly
Difference in (log) road density	-	gROADS	Varying	Static
Gower environmental distance	-	-	-	-



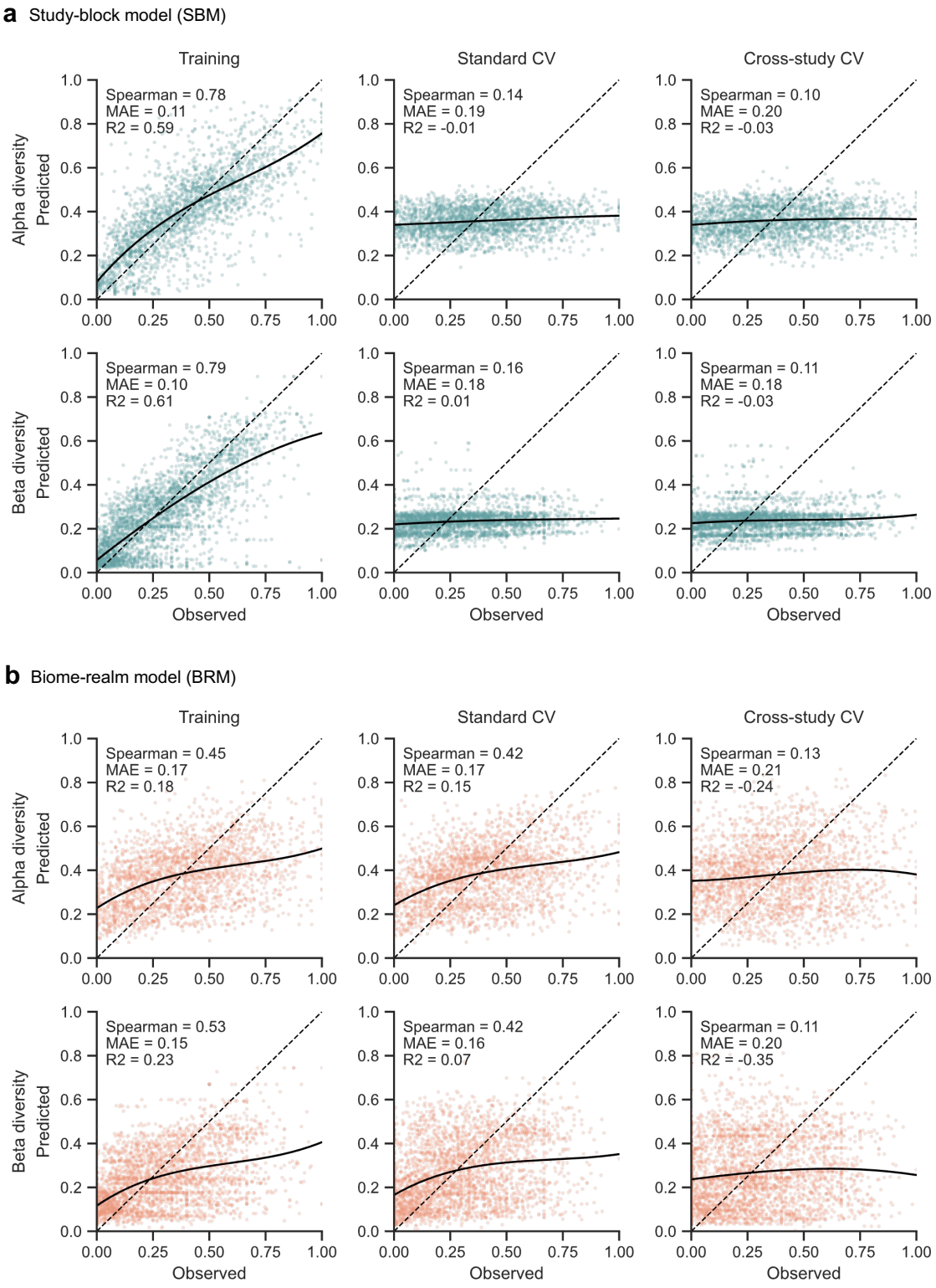
Extended Data Fig. 3 | Response distributions and number of sites per study. a-d, Data distributions of the response variables for the alpha and beta diversity models. Alpha diversity is measured as the geometric mean abundance per site (for the SBM and BRM models, **a**) or per site and taxa (BTM, **c**). Beta diversity is measured as the Bray-Curtis similarity between pairs of sites and reference sites within a study (for the SBM and BRM models, **b**), or between taxonomic groups at pairs of sites (BTM model, **d**). In practice, because of the broad taxonomic grouping used, the difference between the distributions are very small. **e**,**f**, Number of sampling sites per study, all studies (**e**) and filtered to studies with maximum 200 sites (**f**).



Extended Data Fig. 4 | Biome-realm hierarchical groups. Number of source studies and sampling sites (in parenthesis) in each final biome-realm group included in the BRMs, based on the five study and 100 site thresholds. Some groups are at the biome-realm level, while others have been rolled up to the biome level. Groups that were rolled up to the overall global level have been excluded. In practice, the roll-up thresholds are applied for each training fold, but for simplicity we show the results of applying them to the overall data.



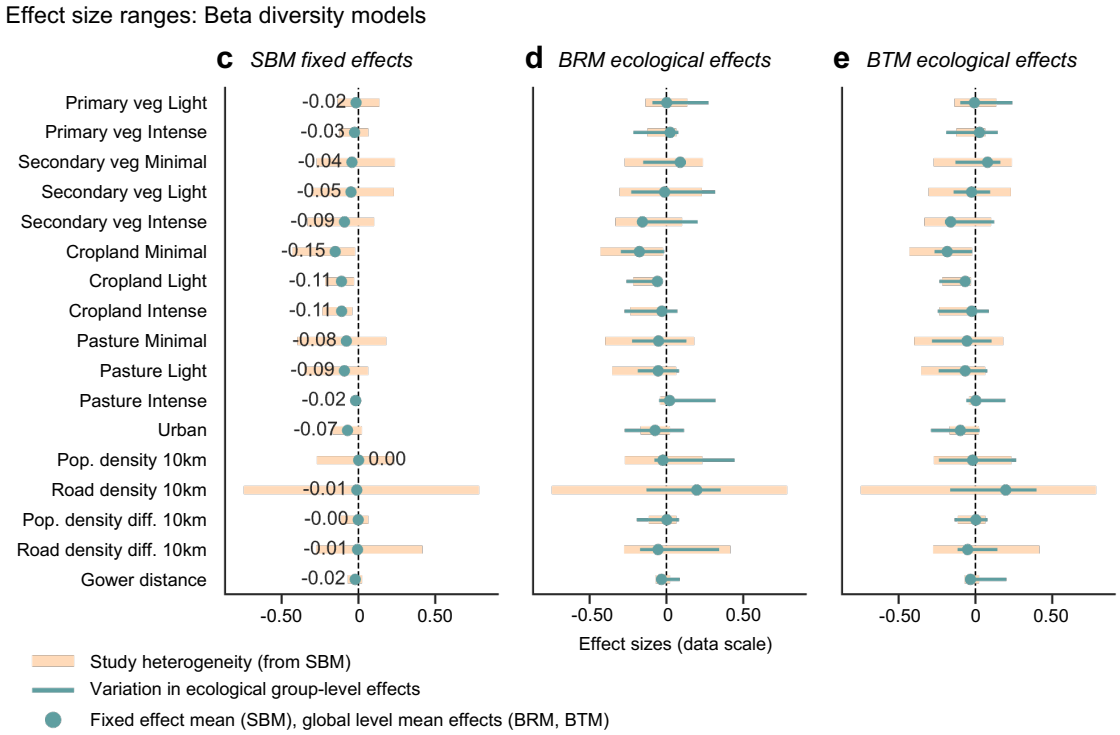
Extended Data Fig. 5 | Biome-taxa hierarchical groups. Number of source studies and sampling sites (in parenthesis) in each final biome-taxa group included in the BTMs, based on the five study and 100 site thresholds. Some groups are at the biome-realm level, while others have been rolled up to the biome level. Groups that were rolled up to the overall global level have been excluded. In practice, the roll-up thresholds are applied for each training fold, but for simplicity we show the results of applying them to the overall data.



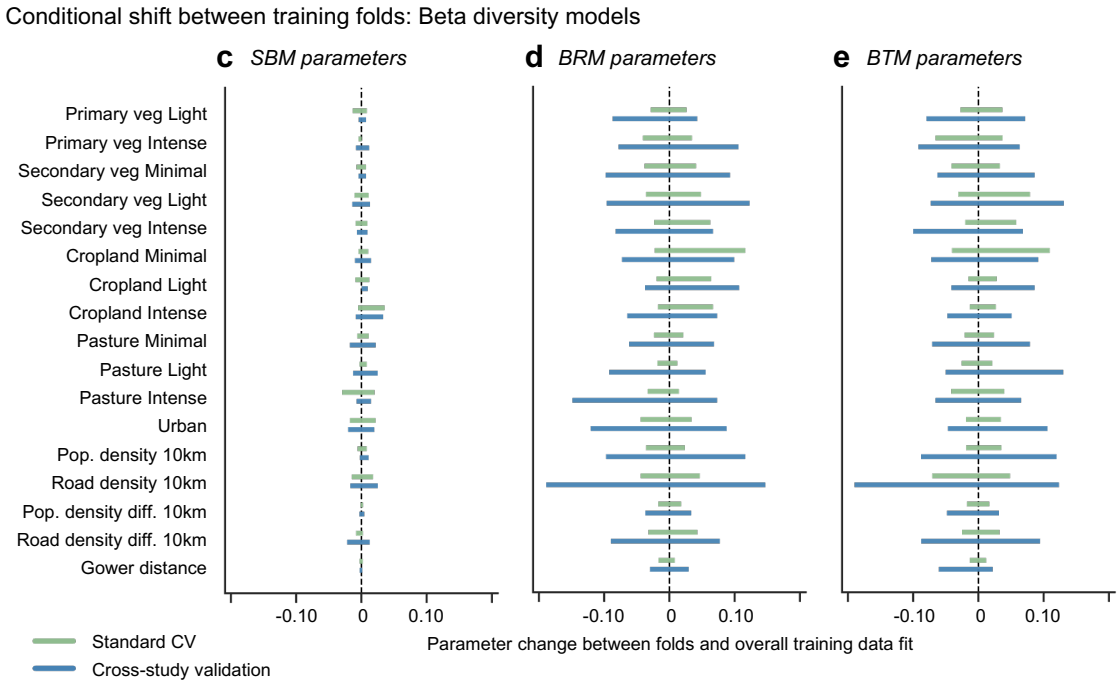
Extended Data Fig. 6 | Model calibration plots. Prediction calibration plots for the SBM (**a**) and BRM (**b**). The first row in each panel shows alpha diversity predictions, the second shows beta diversity predictions. The first column contains predictions on the training data, the second out-of-sample predictions under standard CV, while the third column shows predictions from cross-study validation. The dashed line indicates a perfect predictive fit, and the solid line is a third-degree polynomial fit to the actual predictions. The BTM calibration plots were omitted due to space constraints, but are very similar to the BRM plots.

Extended Data Table 2 | Detailed model performance overview. Model performance across training, standard cross-validation, and cross-study validation. First-row values show the mean, with fold-wise ranges shown beneath.

	Training			Standard CV			Cross-study validation		
Model	ρ	MAE	R^2	ρ	MAE	R^2	ρ	MAE	R^2
SBM alpha	0.78	0.11	0.59	0.14	0.19	-0.01	0.10	0.20	-0.03
	0.78, 0.78	0.11, 0.11	0.59, 0.59	0.10, 0.15	0.19, 0.20	−0.02, 0.00	0.03, 0.21	0.19, 0.20	−0.09, 0.03
BRM alpha	0.45	0.17	0.18	0.42	0.17	0.15	0.13	0.21	-0.24
	0.45, 0.45	0.17, 0.17	0.18, 0.18	0.39, 0.45	0.17, 0.18	0.11, 0.18	0.02, 0.19	0.20, 0.23	−0.51, −0.07
BTM alpha	0.47	0.17	0.20	0.45	0.17	0.17	0.16	0.22	-0.28
	0.47, 0.47	0.17, 0.17	0.20, 0.20	0.44, 0.47	0.17, 0.17	0.15, 0.19	0.11, 0.22	0.20, 0.23	−0.40, −0.18
SBM beta	0.79	0.10	0.61	0.16	0.18	0.01	0.11	0.18	-0.03
	0.79, 0.79	0.10, 0.10	0.61, 0.61	0.13, 0.19	0.17, 0.18	−0.02, 0.02	0.07, 0.35	0.17, 0.19	−0.13, 0.09
BRM beta	0.53	0.15	0.23	0.42	0.16	0.07	0.11	0.20	-0.35
	0.53, 0.53	0.15, 0.15	0.23, 0.23	0.39, 0.44	0.16, 0.17	0.01, 0.12	−0.05, 0.34	0.18, 0.23	−0.79, −0.16
BTM beta	0.57	0.14	0.28	0.52	0.15	0.22	0.16	0.20	-0.33
	0.57, 0.57	0.14, 0.14	0.28, 0.28	0.51, 0.54	0.14, 0.15	0.19, 0.26	−0.06, 0.43	0.16, 0.23	−0.87, 0.08
BRM alpha	0.73	0.13	0.52	0.21	0.20	-0.07	0.10	0.21	-0.25
w/ study controls	0.73, 0.73	0.13, 0.13	0.52, 0.52	0.20, 0.22	0.20, 0.20	−0.10, −0.05	0.00, 0.20	0.20, 0.23	−0.41, −0.14
BTM alpha	0.73	0.13	0.52	0.25	0.20	-0.08	0.09	0.22	-0.30
w/ study controls	0.73, 0.73	0.13, 0.13	0.52, 0.52	0.23, 0.26	0.20, 0.20	−0.11, −0.06	−0.02, 0.22	0.19, 0.25	−0.62, −0.17
BRM beta	0.75	0.11	0.55	0.33	0.17	0.07	0.14	0.19	-0.12
w/ study controls	0.75, 0.75	0.11, 0.11	0.55, 0.55	0.30, 0.36	0.17, 0.18	0.02, 0.11	−0.12, 0.29	0.18, 0.20	−0.62, 0.05
BTM beta	0.76	0.11	0.54	0.38	0.17	0.08	0.18	0.19	-0.15
w/ study controls	0.76, 0.76	0.11, 0.11	0.54, 0.54	0.35, 0.41	0.16, 0.17	0.03, 0.14	0.03, 0.37	0.17, 0.21	−0.42, 0.06



Extended Data Fig. 7 | Effect size ranges, beta diversity models. Estimated ranges of model parameters in the beta diversity models. Yellow bars indicate study heterogeneity, the spread of study-level random effects. These are based on the SBM, and shown in all panels for comparability. Effect sizes are expressed on the scale of the observed data (0–1). **c**, Green circles denote SBM fixed effects, which are averages across all studies. **d**,**e**, Green bars show the range of group-level parameters of the BRM (**d**) and BTM (**e**), respectively. Mean values are left out to not clutter the plots.



Extended Data Fig. 8 | Conditional shift, beta diversity models. Change in estimated model parameters between each training fold fit, and the fitted parameters for the whole dataset, for the beta diversity SBM (**c**), BRM (**d**) and BTM (**e**).