Combining environmental DNA and remote sensing for efficient, fine-scale mapping of arthropod biodiversity

Yuanheng Li[1,2,+], Christian Devenish[3,11,+], Marie I. Tosa[4], Mingjie Luo[1,5], David M. Bell[6], Damon B. Lesmeister[4,6], Paul Greenfield[7,8], Maximilian Pichler[9], Taal Levi[4], and Douglas W. Yu[1,3,10]

[+]Co-first authors

[1]State Key Laboratory of Genetic Resources and Evolution and Yunnan Key Laboratory of Biodiversity and Ecological Security of Gaoligong Mountain, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China 650223

[2]Faculty of Biology, University of Duisburg-Essen, Essen, Germany D-45141

[3]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk, UK NR47TJ

[4]Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State University, Corvallis, Oregon USA 97331

[5]Kunming College of Life Sciences, University of Chinese Academy of Sciences, Kunming, China

[6]Pacific Northwest Research Station, U.S. Department of Agriculture Forest Service, Corvallis, OR, USA 97331

[7]CSIRO Energy, Lindfield, NSW, Australia

[8]School of Biological Sciences, Macquarie University, Australia

[9]Theoretical Ecology, University of Regensburg, Regensburg, Germany

[10]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming Yunnan, China 650223

[11]Current address: School of Geography, Geology and the Environment, Keele University, Staffordshire, ST5 5BG, UK

ABSTRACT

Arthropods contribute importantly to ecosystem functioning but remain understudied. This undermines the validity of conservation decisions. Modern methods are now making arthropods easier to study, since arthropods can be mass-trapped, mass-identified, and semi-mass-quantified into ‘many-row (observation), many-column (species)‘ datasets, with homogeneous error, high resolution, and copious environmental-covariate information. These ‘novel community datasets‘ let us efficiently generate knowledge on arthropod species distributions, conservation values, uncertainty, and the magnitude and direction of human impacts. We use a DNA-based method (barcode mapping) to produce an arthropod-community dataset from 121 Malaise-trap samples, and combine it with 29 remote-imagery layers within a joint species distribution model. With this approach, we generate distribution maps for 76 arthropod species across a 225 km^2 temperate-zone forested landscape. We stack the maps to visualise the fine-scale spatial distributions of species richness, community composition, and site irreplaceability. Old-growth forests show distinct community composition and higher species richness , and stream courses have the highest site-irreplaceability values. By this ‘sideways biodiversity modelling’, we demonstrate the feasibility of biodiversity mapping with sufficient spatial resolution to inform local management choices, while also being efficient enough to scale up to thousands of square kilometres.

Keywords— metabarcoding, environmental DNA, metagenomics, Earth Observation, biodiversity indices, Arthropoda, site irreplaceability, systematic conservation planning, conservation, forestry, machine learning, joint species distribution model

INTRODUCTION

Arthropods contribute in numerous ways to ecosystem functioning (Prather et al., 2013) but are understudied relative to vertebrates and plants (Troudet et al., 2017). This taxonomic bias undermines the validity of conservation decisions when the effects of change in climate, land use, and land cover differ across taxa (Hamilton et al., 2022; Westgate et al., 2014). Also, it is arguable that modern methods now make arthropods *easier* to study than vertebrates and plants, given that arthropods can be mass-trapped and mass-identified (Chua et al., 2023; van Klink et al., 2022). Another logistical advantage is that arthropod community structure is correlated with vegetation structure (Lewinsohn and Roslin, 2008; Zhang et al., 2016), and since vegetation can be measured remotely at large spatial scale via airborne and spaceborne sensors (Bush et al., 2017), remote imagery could also provide large-spatial-scale information on arthropods. In fact, it is already known that spaceborne SAR (synthetic aperture radar) and airborne lidar (Light Detection And Ranging) imagery of fine-scale forest structure can predict the distributions of entomofauna and avifauna (Bae et al., 2019; Müller et al., 2009; Müller and Brandl, 2009; Rhodes et al., 2022).

Successful governance of the biodiversity commons

Arthropod conservation should be seen in the wider context of efficient biodiversity governance. Dietz et al.'s (2003) framework for the successful governance of public goods can be usefully summarised into five elements: (1) knowledge generation, (2) infrastructure provision, (3) political bargaining, (4) enforcement, and (5) institutional redesign. Dietz et al.'s knowledge-generation element asks engineers and scientists to generate *high-quality*, *granular*, *timely*, *trustworthy*, and *understandable* knowledge on ecosystem status and change, values, uncertainty levels, and the magnitude and direction of human impacts.

However, to our knowledge, there is no example of the five elements comprehensively working together to achieve *multi-species* conservation, in large part because the tools, study designs, and analyses needed to generate knowledge on many species at once are complex. This complexity is a barrier to uptake, delaying the institutional redesigns that could operationalise, finance, and scale-up conservation. (See Supplementary Information: "Dietz's five elements" for an example of the five elements working together to create a single-species biodiversity-offset market).

Our focus in this study is therefore to demonstrate how to efficiently generate *high-quality*, *granular*, *timely*, *trustworthy*, and *understandable* knowledge on status and change in arthropod biodiversity, conservation value, uncertainty levels, and the magnitude and direction of human impacts.

We use the management of National Forests in the United States as our test case for multi-species biodiversity conservation. This management should follow the doctrine outlined in the 1960 Multiple-Use Sustained-Yield Act that requires management and utilisation of natural resources to satisfy multiple competing interests and to maintain the natural resources in perpetuity (Carter et al., 2019; Hobbs et al., 2010; Loomis, 2002). Although US law mandates that each use be given equal priority, implementation is stymied by a lack of biodiversity data such as distribution maps of large numbers of species to identify areas of high conservation value that can be protected while still supporting extractive uses in other areas. Moreover, the species distribution maps should be regularly updated so that the impacts of management interventions can be inferred, feeding back to adaptive management (Frankham, 2010; Bush et al., 2017).

High-throughput arthropod inventories

Now though, there are new technologies capable of efficiently and granularly capturing biodiversity information, via DNA isolated from environmental samples (eDNA) and via electronic sensors (bioacoustics, cameras, radar) (Besson et al., 2022; Bohmann et al., 2014; Bush et al., 2017; Christin et al., 2019; Pawlowski et al., 2020; Ruppert et al., 2019; Tosa et al., 2021; van Klink et al., 2022; Chua et al., 2023). Many of these methods start with DNA-based taxonomic assignment (‘DNA barcoding’ Hebert et al., 2003) and vary in how the DNA is collected and processed. For instance, large numbers of arthropods can efficiently be individually DNA-extracted and sequenced to produce count datasets (Ratnasingham, 2019; Srivathsan et al., 2021). These DNA-barcoded specimens (plus human-identified specimens) can optionally be used to annotate specimen images to train deep-learning models to scale up identifications (Chua et al., 2023; van Klink et al., 2022). Alternatively, DNA from arthropods can be extracted *en masse* from traps (Ji et al., 2013) or from environmental substrates, such as water washes of flowers (e.g. Thomsen and Sigsgaard, 2019) and mass-sequenced. These latter processing pipelines are known as ‘metabarcoding’ or ‘metagenomics’, depending on whether the target DNA-barcode sequence is PCR-amplified (both described in Bush et al., 2017).

All these methods produce ‘novel community data’, which Hartig et al. (2023) describe as ‘many-row (observation), many-column (species)’ datasets, therefore making possible high spatial and/or temporal resolution and extent. Novel community data contain some form of abundance information, ranging from counts to within-species abundance change (Luo et al., 2023; Diana et al., 2022) to presence *and* absence, and because the methods are automated and standardised, the errors in these datasets tend to be homogeneous (e.g. minimal observer effects), which facilitates their correction given appropriate sample replicates and statistical models.

‘Sideways’ biodiversity modeling and site irreplaceability ranking

It is natural to think about combining novel community data with copious environmental-covariate information in the form of continuous-space remote-imagery layers (and/or with continuous-time acoustic time series) to produce continuous spatio(-temporal) biodiversity data products (Bush et al., 2017; He et al., 2015; Kwok, 2018; Leitão and Santos, 2019; Lin et al., 2021; Pettorelli et al., 2018; Cavender-Bares et al., 2022; Hartig et al., 2023). Here we do just this, combining a point-sample dataset of Malaise-trapped arthropods with continuous-space Landsat and lidar imagery within a joint species distribution model (JSDM) framework (JSDM Ovaskainen and Abrego, 2020; Pichler and Hartig, 2021; Warton et al., 2015). We were able to produce distribution maps for 76 arthropod species across a forested landscape. Because this landscape is characterised by overlapping gradients of environmental conditions (e.g. elevation, distance from streams and roads) and mosaics of management (e.g. clearcuts, old-growth), we can estimate the effects of different combinations of natural and anthropogenic drivers on arthropod biodiversity, including combinations that were not included in our sample set. We can also subdivide the landscape into management units and rank them by conservation value, to inform decision-making in this multi-use landscape.

The above approach is a direct test of a protocol originally proposed by Bush et al. (2017) and more formally described by Pollock et al. (2020) under the name ‘sideways’ biodiversity modelling. In short, sideways biodiversity models (1) integrate “the largely independent fields of biodiversity modeling and conservation” and (2) include large numbers of species in conservation planning instead of using habitat-based metrics. Or in plain language, we use remote-sensing imagery to fill in the blanks between our sampling points, which creates a continuous map of arthropod biodiversity that we can use to study arthropod ecology and guide conservation.

MATERIALS AND METHODS

In short, we combine DNA-based species detections, remote-sensing-derived environmental predictors, and joint species distribution modelling to predict and visualise the fine-scale distribution of arthropods across a large forested landscape. We the stack the individual species distributions to map species richness, compositional distinctiveness, and conservation value across the landscape. For the detailed protocol and explanations of the field, laboratory, bioinformatic, and statistical methods, see Supplementary Information: Materials and Methods.

Model Inputs

Field data collection

We collected with 121 Malaise-trap samples for seven days into 100% ethanol at 89 sampling points in and around the H.J. Andrews Experimental Forest (HJA), Oregon, USA in July 2018 (Figure 1). Sites were stratified by elevation, time since disturbance, and inside and outside the HJA (inside: a long-term research site with no logging since 1989; outside: continued active management). HJA represents a range of previously logged to primary forest, but with notably larger areas of mature and old-growth forest reserves than the regional forest mosaic, which consists of short-rotation plantation forests on private land and a recent history of active management on public land.

Wet-lab pipeline and bioinformatics

DNA extraction and sequencing

We extracted the DNA from each Malaise-trap sample by soaking the arthropods in a lysis buffer and sent it to Novogene (Beijing, China) for whole-genome shotgun sequencing.

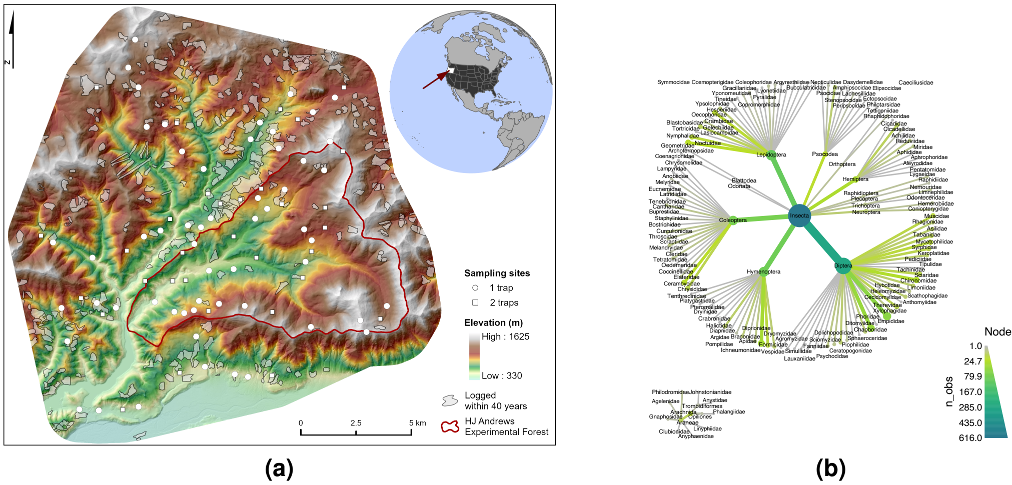

N
Sampling sites
○ 1 trap
□ 2 traps
Elevation (m)
High : 1625
Low : 330
Logged within 40 years
HJ Andrews Experimental Forest
0
2.5
5 km
(a)
Insecta
Lepidoptera
Psocodea
Orthoptera
Hemiptera
Raphidioptera
Plecoptera
Trichoptera
Neuroptera
Coleoptera
Blattodea
Odonata
Hymenoptera
Diptera
Symmocidae Cosmopterigidae Coleophoridae Argyresthiidae Nepticulidae Dasydemellidae Caeciliusidae
Gracillariidae Lyonetiidae Bucculatricidae Amphipsocidae Elipsocidae
Yponomeutidae Pyralidae Psocidae Lachesillidae
Tineidae Copromorphidae Ectopsocidae
Ypsolophidae Stenopsocidae Peripsocidae Philotarsidae
Hesperiidae Tettigoniidae
Oecophoridae Rhaphidophoridae
Blastobasidae Crambidae Cicadidae
Tortricidae Gelechiidae Cicadellidae
Lasiocampidae Achilidae
Nymphalidae Noctuidae Reduviidae
Geometridae Miridae
Archotermopsidae Aphididae
Coenagrionidae Aphrophoridae
Chrysomelidae Aleyrodidae
Lampyridae Pentatomidae
Anobiidae Lygaeidae
Melyridae Raphidiidae
Eucnemidae Nemouridae
Latridiidae Limnephilidae
Tenebrionidae Odontoceridae
Cantharidae Hemerobiidae
Buprestidae Coniopterygidae
Staphylinidae Muscidae
Bostrichidae Rhagionidae
Curculionidae Asilidae
Throscidae Tabanidae
Scraptiidae Mycetophilidae
Melandryidae Syrphidae
Cleridae Keroplatidae
Tetratomidae Pediciidae
Oedemeridae Tipulidae
Coccinellidae Tachinidae
Elateridae Sciaridae
Cerambycidae Chironomidae
Chrysididae Hybotidae Limoniidae
Tenthredinidae Heleomyzidae Scathophagidae
Platygastridae Cecidomyiidae Anthomyiidae
Pteromalidae Therevidae
Dryinidae Xylophagidae
Crabronidae Phoridae
Halictidae Ditomyiidae Empididae
Diapriidae Diprionidae Chaoboridae
Argidae Braconidae Dryomyzidae Dolichopodidae
Apidae Agromyzidae Sciomyzidae Sphaeroceridae
Pompilidae Formicidae Fanniidae Piophilidae
Ichneumonidae Vespidae Simuliidae Ceratopogonidae
Lauxaniidae Psychodidae
Philodromidae Johnstonianidae
Agelenidae Anystidae
Trombidiformes
Arachnida Phalangiidae
Gnaphosidae Opiliones
Araneae
Linyphiidae
Clubionidae
Anyphaenidae
Nodes
n_obs
1.0
24.7
79.9
167.0
285.0
435.0
616.0
(b)

Figure 1. Sampling design and taxonomic diversity of the Malaise trapping campaign. (a). Sampling points in and around the H.J. Andrews Experimental Forest (red line), Oregon, USA. The study area consists of old-growth and logged (gray patches) deciduous and evergreen forest under different management regimes. Arthropods were sampled with Malaise traps at 89 sampling points in July 2018, with one trap at 57 points (white circles) and with two traps 40 m apart at 32 points (white squares). Elevation indicated with a green to white false-color gradient. (b). Taxonomic distribution of 190 Operational Taxonomic Units (OTUs) from the samples. Node size and color are scaled to the number of OTUs.

Creating a barcode reference database using Kelpie in-silico PCR

We used 'in-silico PCR' implemented in `Kelpie` (Greenfield et al., 2019) to find DNA barcodes in the shotgun-sequence datasets. `Kelpie` searches for reads that match the two ends of the DNA barcode region and then searches for overlapping reads, ultimately assembling whole barcodes. We used the BF3+BR2 primers from Elbrecht et al. (2019), which bookend a 418-bp fragment of the COI DNA barcode. `Kelpie` assembled 5560 unique DNA-barcode sequences, and after a series of bioinformatic 'denoising' steps to remove all sequence variation within species, we were left with a reference barcode dataset of 1225 unique sequences, each of which potentially represents a species and is thus known as an "operational taxonomic unit" or OTU.

Read mapping to reference barcodes

We then mapped the individual shotgun reads of each sample to the reference barcodes, creating a 121-sample × 1225-OTU table. A species was accepted as being in a sample if its reads mapped at high quality along more than 50% of its barcode length, following acceptance criteria from Ji et al. (2020).

Environmental covariates

To predict species occurrences in the areas between the sampling points, we collected 58 continuous-space predictors (Table 1S), relating to forest structure, vegetation reflectance and phenology, topography, and anthropogenic features, restricting ourselves to predictors that can be measured remotely. The forest-structure variables were from airborne lidar data collected from 2008 to 2016, which correlate with forest structure in US Pacific Northwest coniferous forests, such as mean diameter, canopy cover, and tree density (Kane et al., 2010). The vegetation-related variables came from Landsat 8 individual bands, plus standard deviation, median, 5% and 95% percentiles of those bands over the year, and indices of vegetation status e.g. Normalized Difference Vegetation Index (NDVI). Both the proportion of canopy cover and annual Landsat metrics were calculated within radii of 100, 250 and 500 m, given that vegetation structure at different spatial scales is known to drive arthropod biodiversity (Müller et al., 2014). The topography variables were calculated from lidar ground returns, including elevation, slope, Eastness and Northness split from aspect, Topographic Position Index (TPI), Topographic Roughness Index (TRI) (Wilson et al., 2007), Topographic Wetness Index (TWI) (Metcalfe et al., 2018), and distance to streams, based on a vector stream network (`http://oregonexplorer.info`, accessed 24 Oct 2019). The anthropogenic variables include distance to nearest road, proportion of area logged within the last 100 and 40 years within radii of 250, 500, and 1000 m, and a categorical variable of inside or outside the boundary of the H.J. Andrews Experimental Forest. They are not directly derived from remote-sensing data, but we included them because they could be derived from remote-sensing imagery. We then reduced our 58 environmental covariates to 29, removing the covariates that were most correlated with the others (as measured by Variance Inflation Factor). The 29 retained covariates include six anthropogenic activities, two raw Landsat bands, seven indices based on annual Landsat data, six canopy/vegetation related variables from LiDAR, and eight topography variables (Table 1S, 3S), which we mapped across the study area at 30 m resolution.

Statistical Analyses

Species inputs

We converted the sample × species table to presence-absence data ($1/0$), and we only included species present at ≥ 6 sampling sites across the 121 samples. Our species dataset was thus reduced to 190 species in two classes, Insecta and Arachnida (Figure 1b).

Joint Species Distribution Model

The general idea behind species distribution modelling is to ‘predict a species’ distribution’, using the species’ observed incidences and the environmental-covariate values at those points, to ‘fit’ a model that predicts the former from the latter. After model fitting, the species’ probability of presence is predicted over the rest of the sampling area, where environmental conditions are known but species’ incidences are not.

Tuning and testing

The statistical challenge is to avoid overfitting, which is when the fitted model does a good job of predicting the species' incidences at the sampling points that were used to fit the model in the first place but does a bad job of predicting the species over the rest of the landscape. Overfitting is likely in our dataset because many of our species are rare, there are many candidate remote-sensing covariates, and we expect that any relationships between remote-sensing-derived covariates and arthropod incidences are indirect and thus complex, necessitating the use of flexible mathematical functions.

To minimise overfitting, we used regularisation and five-fold cross-validation. Regularisation uses penalty terms during model fitting to favour a relatively simple set of covariates, and cross validation finds the best values for those penalty terms. First, we randomly split the species incidence data from the 121 samples in 89 sampling points into 75% training data ($n = 91$) and 25% test data ($n = 30$) (Figure 1M). The training data was used to try 1000 different hyperparameter combinations, some of which are the penalty terms, to find the combination that achieves the highest predictive performance on the training data itself (Figure 1M). The model with this combination was then applied to the 25% test data to measure true predictive performance. To fit the model, we used the joint species distribution modelling `R` package `sjSDM 1.0.5` (Pichler and Hartig, 2021), with the DNN (deep neural net) option to to account for complex, nonlinear effects of environmental covariates, which suits our dataset of many species with few data points and many covariates.

Variable importance with explainable-AI (xAI)

The mathematical functions used in neural net models are unknown, but it would be useful to identify the covariates that contribute the most to explaining each species incidences. We therefore carried out an 'explainable-AI' (xAI) analysis, using the `R` package `flashlight 0.8.0` (Mayer, 2021). In short, for each environmental-covariate, we shuffled its values in the dataset and estimated the drop in explanatory performance on the training data. The most important covariate is the one that, when permuted, degrades explanatory performance the most.

Prediction and visualisation of species distributions

Finally, after applying the final model to the test dataset, we identified 76 species that had moderate to high predictive performance (AUC $\geq$ 70%). We used the fitted model and the environmental-covariates to predict the probability of each species' incidence in each grid cell of the study area ('filling in the blanks' between the sampling points). The output is 76 individual and continuous species distribution maps, which we stacked to carry out three landscape analyses. First, we counted the number of species predicted to be present (probability of presence $\geq$ 50%) in each grid square to produce a species richness map. Second, we carried out a dimension-reduction analysis, also known as ordination, using the T-SNE method (van der Maaten and Hinton, 2008; Krijthe, 2015) to summarise species compositional change across the landscape. Pixels that have similar species compositions receive similar T-SNE values, which can be visualised. Third, we calculated Baisero et al.'s (2022) site-irreplaceability index for every pixel. This index is the probability that loss of that pixel would prevent achieving the conservation target for at least one of the 76 species, where the conservation target is set to be 50% of the species' total incidence.

Finally, we carried out *post-hoc* analyses by plotting site irreplaceability, composition (T-SNE), and species richness against elevation, old-growth structural index (Davis et al., 2015), and inside/outside HJA.

RESULTS

Model Inputs

DNA/Taxonomic data

The 121 samples from July and August 2018 were sequenced to a mean depth of 29.0 million read-pairs 150-bp (median 28.9 M, range 20.8-47.1 M), of which we used only the July samples. Of the 190 OTUs used in our joint species distribution model, 183 were assigned to Insecta, and 7 to Arachnida (Figure 1b). All OTUs could be assigned to order level, 178 to family level, 131 to genus level, and 66 to species level (Figures 1b, 2S).

Statistical Analyses

Model performance and xAI

Across all species together, the final sjSDM model achieves median and mean explanatory-performance values of $AUC = 0.86$ and 0.86, respectively, where the AUC (Area Under the Curve) metric equals 1 for a model with 100% correct predictions and 0 for 100% incorrect predictions. The model's median and mean predictive AUC (i.e. on the test data) are 0.67 and 0.67 (Figure 2Ma). Predictive AUC is a measure of model generality, and the fact that explanatory AUCs are greater than predictive AUCs demonstrates how fitting a model to a particular dataset results in a degree of overfitting. Per species, mean AUC values range from 0 (fail completely) to 1 (predict perfectly), and this variation was not explained by by species' taxonomic family or prevalence (% presence in sampling points).

Out of 29 environmental covariates, 18 (Table 1S) were the most important for at least one species (Figure 2Mb). Elevation and Topographic Roughness Index (TRI) were the most important covariates for the most species. Eleven environmental covariates were the most important for at least one species in terms of interaction effects of the variables, with elevation and TRI again being the most important (Figure 6S).

Prediction and visualization of species distributions

Finally, we reduced the dataset to the 76 species with individual predictive AUCs > 0.7 (mean $= 0.834$), and for each, we generated individual distribution maps across the study area, which differ in amount and distribution of the areas with high predicted habitat suitability (Figures 2 E-L, 7S). We then stacked the maps to estimate the fine-scale spatial distributions of species richness, community composition, and site irreplaceability across the study area (Figure 2). Site irreplaceability, which is a core concept in systematic conservation planning, ranks each site by its importance to the "efficient achievement of conservation objectives" (Kukkala and Moilanen, 2013). In practice, high-irreplaceability sites tend to house many species with small ranges and/or species with large ranges that we wish to conserve a large fraction of, such as endangered species.

Greater species richness was predicted for areas without recent logging, especially within the northeast and southeast sectors of the H.J. Andrews Experimental Forest (HJA), on west-facing slopes, and in the south of the study area (Figure 2 A). A *post-hoc* analysis found a non-linear increase in species richness in the largest patches of old-growth forest, which are inside the HJA (Figure 3 A, B).

T-SNE ordination reveals spatial patterning in species composition (Figure 2 C, D). T-SNE-1 is clearly correlated with elevation (compare Figures 1a and 3 C), whereas T-SNE-2 appears to be correlated with the extent of surrounding old-growth forest, at middle elevations (Figure 3 C). Finally, site irreplaceability clearly follows stream courses, which are mostly at low elevations (Figure 2 B) and cover a small portion of the total landscape. As a result, *post-hoc* analysis also shows that irreplaceability decreases with elevation but finds no relationship between irreplaceability and surrounding old-growth forest (Figure 3 D).

DISCUSSION

We combined *in-silico* barcode-mapping data derived from 121 arthropod bulk samples in 89 sampling points spread over a 225 km^2 working and primary forest with 29 environmental covariates (Figure 3S) from Landsat, lidar, and other layers that covered information on forest structure, vegetation condition, topography, and anthropogenic impact. We used a joint species distribution model with a neural net to predict the fine-scale spatial distributions of 76 Insecta and Arachnida species with a high degree of estimated predictive performance (all individual predictive AUCs > 0.7, mean $= 0.834$) (Figure 2Ma). The model made good use of the 29 environmental covariates, with 18 of them being the most important for at least one species (Figure 2Mb), with elevation and Topographic Roughness Index (TRI) most important covariates for the most species. These two covariates were also the most frequently most important in terms of their interactions with other covariates (Figure 6S).

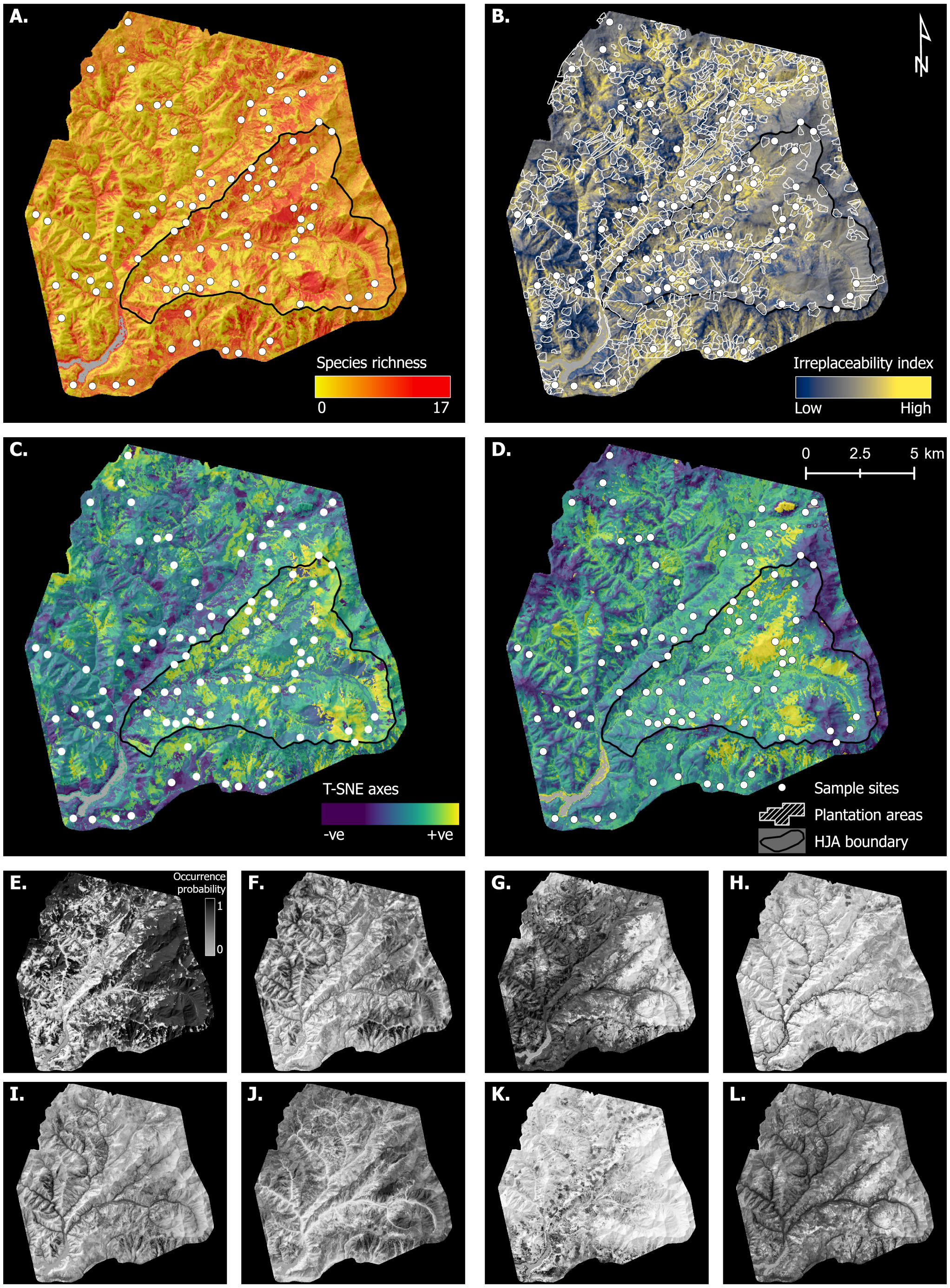


Figure 2. JSDM-interpolated spatial variation in species richness, irreplaceability, and composition, plus examples of individual species distributions. A. Species richness. B. Site beta irreplaceability, showing areas of forest plantation. C-D. T-SNE axes 1 and 2. White circles indicate sampling points, white polygons indicate plantation areas (i.e. a record of logging in the last 100 years), and the black triangle delimits the H.J. Andrews Experimental Forest (HJA, Fig. 1). E-L. Selected individual species distributions (all species in Figure 7S), with BOLD ID, predictive AUC, and prevalence. E. Rhagionidae gen. sp. (`BOLD: ACX1094`, AUC: 0.91, Prev: 0.64). F. *Plagodis pulveraria* (`BOLD: AAA6013`, AUC: 0.81, Prev: 0.23). G. *Phaonia* sp.(`BOLD: ACI3443`, AUC: 0.80, Prev: 0.65). H. *Melanostoma mellinum* (`BOLD: AAB2866`, AUC: 0.90, Prev: 0.11). I. *Helina* sp. (`BOLD: ACE8833`, AUC: 0.73, Prev: 0.23). J. *Bombus sitkensis* (`BOLD: AAI4757`, AUC: 0.98, Prev: 0.23). K. *Blastobasis glandulella* (`Bold: AAG8588`, AUC: 0.86, Prev: 0.18). L. *Gamepenthes* sp. (`BOLD: ACI5218`, AUC: 0.77, Prev: 0.57).

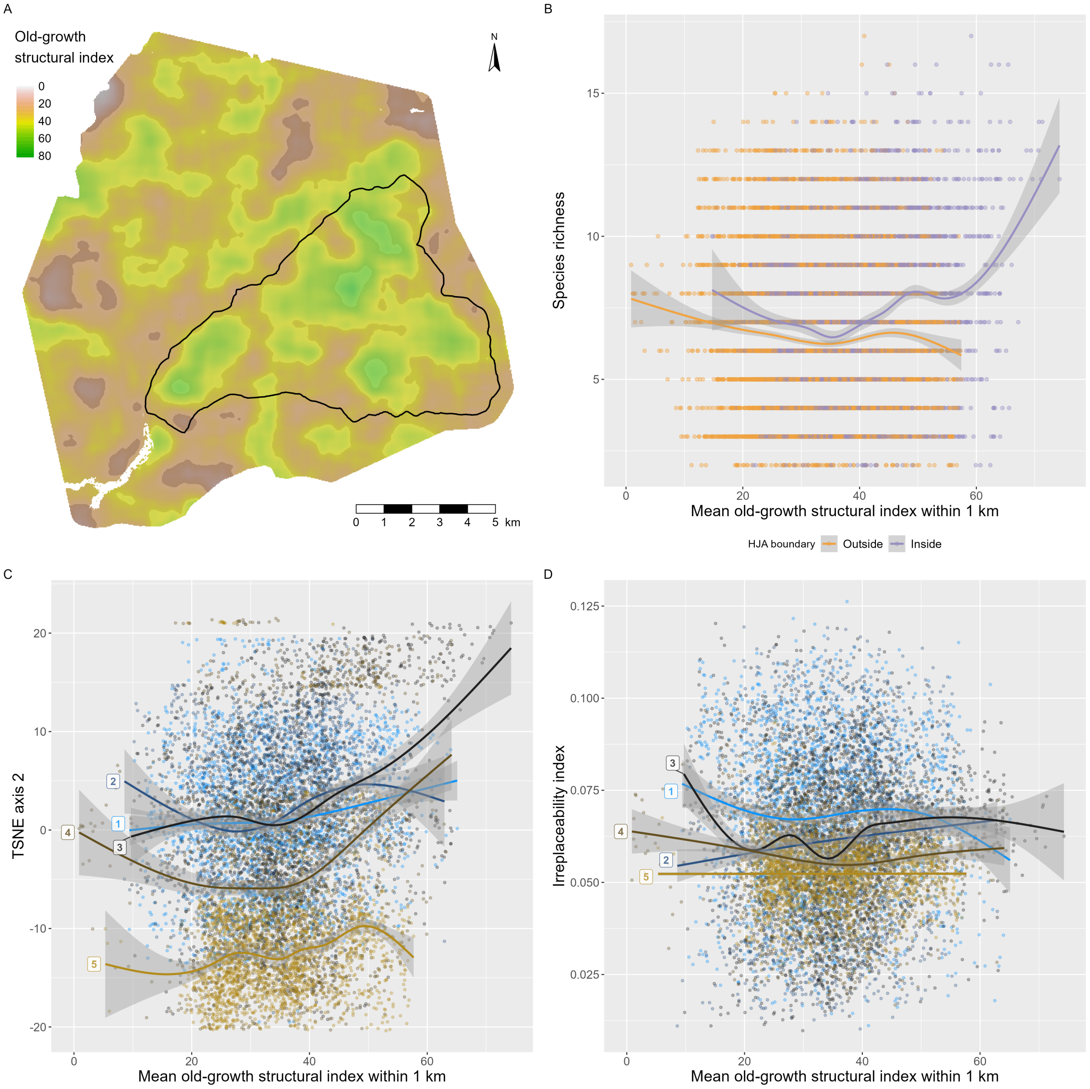


Figure 3. *Post-hoc* analysis of species richness, composition, and irreplaceability patterns in Figure 2, in relation to old-growth structural index (OGSI). A. Smoothed OGSI, showing principal patches of old-growth forest inside and outside the H.J. Andrews Experimental Forest (black triangle). The HJA has the largest patches of old-growth forest. B. Species richness increases in the parts of the HJA with the highest OGSI values (compare with Figure 2 A). C. Species compositions in the largest old-growth patches, which are at elevation bands 3 and 4, are distinct from the rest of the landscape (compare with Figure 2 D). D. Irreplaceability shows no relationship with OGSI at any elevation (compare with Figure 2 B). Elevation bands (blue to brown colour gradient) 1: 380 – 620; 2: > 620 – 865; 3: > 865 – 1115; 4: > 1115 – 1365; 5: > 1365 – 1615 m above sea level. Splines fit using `mgcv` (Wood, 2017).

By stacking the individual maps, we created *granular* maps of arthropod biodiversity metrics across our study area: species richness, community composition, and site irreplaceability (Figure 2). After viewing these maps, we observed *post-hoc* that species richness is higher and that species composition is distinct in the largest patches of old-growth forest (Figure 3 B, C), but not exclusively so. Irreplaceability, as we have defined it here using Baisero et al.'s (2022) formulation, which does not take connectivity or ecosystem functions into account, is highest along stream courses (Figure 3 D), which are dominated by species with high occurrence probabilities covering a small area (Figure 7S). Irreplaceability is not higher in old-growth forest, given that old-growth is not a rare habitat in our study area. We consider the patterns observed in (Figure 3) to be hypotheses for future testing, and thus we do not calculate statistical significance values.

A biodiversity map is more *understandable* than is an analysis of data points and can be compared directly with land-use maps. In principle, these datasets and products can also be *timely*, given that the creation of DNA-based datasets can be outsourced to commercial labs in some countries with turnaround times measured in weeks. Information *quality* can be assessed via prediction performance 2Ma, and even *trustworthiness* can be assessed via a combination of proof-of-work GPS surveyor tracking and independent re-sampling, given that sampling is standardized (Hartig et al., 2023).

In summary, we show how to generate information on arthropod spatial distributions with a high-enough resolution to make it useful and understandable for local management while also being efficient and standardised enough to scale up to thousands of square kilometres. In Supplementary Information: Bioinformatic and Statistical Analysis Methods, we discuss four methodological caveats (Irreplaceability calculations, false-negative error, errors in environmental covariates, and choice of JSDM software and interpretation). We conclude by briefly reviewing potential applications of this approach.

Potential applications of efficient, fine-scale, and large-scale species distribution mapping

This study demonstrates how the major steps of species distribution mapping are enjoying major efficiency gains (Besson et al., 2022; Bush et al., 2017; Speaker et al., 2022; Tosa et al., 2021). Large numbers of point samples can be characterized to species resolution via DNA sequencing and/or electronic sensors, large numbers of environmental covariates are available from near- and remote-sensing sources (Lock et al., 2022), and GPU-accelerated deep learning algorithms can be used to both accelerate and improve model fitting on these larger datasets (Pichler and Hartig, 2021, 2023). Although this study focused on arthropods, a wide range of animal, fungal, and plant taxa can be detected using DNA extracted from water, air, invertebrate, and soil samples (Abrego et al., 2018; Bohmann et al., 2014; Guimarães Sales et al., 2019; Ji et al., 2022; Leempoel et al., 2019; Lin et al., 2021; Massey et al., 2022; Rodgers et al., 2017; Thomsen and Sigsgaard, 2019; Tilker et al., 2020), with river networks being an especially promising way to scale up sampling over large areas (Guimarães Sales et al., 2019; Lyet et al., 2021).

As a result, it is possible to envisage implementing Pollock et al.'s (2020) vision of using 'sideways' species-based biodiversity monitoring to subdivide whole landscapes for ranking by conservation value (see also Cavender-Bares et al., 2022). One potential benefit would be to interpret remote-sensing imagery in terms of species compositions, thus improving the efficiency of habitat-based offset schemes, such as England's Biodiversity Net Gain legislation, which has been criticized for undervaluing some habitat types, such as scrubland, that are known to support high insect diversity and abundance (Weston, 2021).

Recent studies have also shown that timely and/or fine-resolution biodiversity distribution data can potentially improve conservation decision-making, over that informed by historical distribution data. Ji et al. (2022) used 30,000 leeches mass-collected by park rangers to map for the first time the distributions of 86 species of mammals, amphibians, birds and squamates across a 677 km^2 nature reserve in China, finding that domestic species (cows, goats, and sheep) dominated at low elevations, whereas most wildlife species were limited to mid- and high-elevation portions of the reserve. Before this study, no comprehensive survey had taken place since 1985, impeding assessment of the reserve's effectiveness, which is a general problem in the management of protected areas (Maxwell et al., 2020). Chiaverini et al. (2022) used camera-trap data to extrapolate the distributions of vertebrate species richness across Borneo and Sumatra and found that high species richness areas did not correlate well with IUCN range maps, which are based on historical distribution data (`https://www.iucnredlist.org`, accessed 18 April 2022). Finally, Hamilton et al. (2022) compiled decades of standardized biodiversity inventory data for 2216 species in the United States and extrapolated across the country (excluding Alaska and Hawaii) to identify areas of unprotected biodiversity importance (using a measure similar in spirit to site irreplaceability, i.e. protection-weighted range-size rarity). Because the resulting maps were *granular* (990 m), Hamilton et al. (2022) were able to compare species distributions with land tenure data, including protected areas, and found large concentrations of unprotected species in areas not previously flagged in continental- and regional-scale analyses, in part due to the inclusion of taxa not normally included in such analyses (especially plants, freshwater invertebrates, and pollinators).

Conclusion

A major difficulty for basic and applied community ecology is the collection of many standardised observations of many species. DNA-based methods provide capacity for collecting data on many species at once, but costs scale with sample number. In contrast, remote-sensing imagery provides continuous-space and near-continuous-time environmental data, but most species are invisible to electronic sensors. By combining the two, we show that it is possible to create a combined spatio(temporal) data product that can be interrogated in the same way as an exhaustive community inventory.

ACKNOWLEDGMENTS

We thank field technicians BP Murley, SD Sparrow, and ME Yates. DWY and ML were supported by the Key Research Program of Frontier Sciences, CAS (QYZDY-SSW-SMC024), the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDA20050202, the State Key Laboratory of Genetic Resources and Evolution (GREKF19-01, GREKF20-01, GREKF21-01) at the Kunming Institute of Zoology, and the University of Chinese Academy of Sciences. DWY was also supported by the University of East Anglia and a Leverhulme Trust Research Fellowship (RF-2017-342), and benefited from the sCom Working Group at iDiv.de. MIT was supported by the National Science Foundation-funded H.J. Andrews Long-Term Ecological Research (LTER) program (DEB-1440409), Oregon State University, and the ARCS Oregon Chapter. Field data collection was funded by Oregon State University, the Pacific Northwest Research Station, and the U.S. Department of Agriculture Forest Service. Lidar data processing was supported by the National Science Foundation-funded H. J. Andrews LTER Program (DEB-2025755, DEB-1440409) and the Pacific Northwest Research Station. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official U.S. Department of Agriculture or U.S. Government determination or policy. The use of trade or firm names in this publication is for reader information and does not imply endorsement by the U.S. Government of any product or service.

AUTHORS' CONTRIBUTIONS

DWY and TL conceived the project. TL, MIT, DMB, DBL designed the sampling methodology; MIT and ML collected the data; YL, CD, ML, and DWY analyzed the data; PG and MP contributed unpublished software; YL, CD, and DY led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY

Raw sequence data are archived at NCBI Short Read Archive BioProject PRJNA869351 Reviewer link until 1 Sep 2023, afterwards public. All scripts and data tables (from bioinformatic processing to statistical analysis to figure generation) are available at `https://github.com/chnpenny/HJA_analyses_Kelpie_clean/releases/tag/v1.1.0` and archived at doi:10.5281/zenodo.8303158.

COMPETING INTERESTS

DWY is a co-founder of NatureMetrics (www.naturemetrics.com), which provides commercial metabarcoding services. All other authors have no competing interests.

REFERENCES

Abrego, N., Norros, V., Halme, P., Somervuo, P., Ali-Kovero, H., and Ovaskainen, O. (2018). Give me a sample of air and I will tell which species are found from your region: Molecular identification of fungi from airborne spore samples. *Molecular Ecology Resources*, 18(3):511–524.

Bae, S., Levick, S. R., Heidrich, L., Magdon, P., Leutner, B. F., Wöllauer, S., Serebryanyk, A., Nauss, T., Krzystek, P., Gossner, M. M., Schall, P., Heibl, C., Bässler, C., Doerfler, I., Schulze, E.-D., Krah, F.-S., Culmsee, H., Jung, K., Heurich, M., Fischer, M., Seibold, S., Thorn, S., Gerlach, T., Hothorn, T., Weisser, W. W., and Müller, J. (2019). Radar vision in the mapping of forest biodiversity from space. *Nature Communications*, 10(1):4757.

Baisero, D., Schuster, R., and Plumptre, A. J. (2022). Redefining and mapping global irreplaceability. *Conservation Biology*, 36(2):e13806.

Besson, M., Alison, J., Bjerge, K., Gorochowski, T. E., Høye, T. T., Jucker, T., Mann, H. M. R., and Clements, C. F. (2022). Towards the fully automated monitoring of ecological communities. *Ecology Letters*, 25(12):2753–2775.

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., and de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, 29(6):358–367.

Bush, A., Sollmann, R., Wilting, A., Bohmann, K., Cole, B., Balzter, H., Martius, C., Zlinszky, A., Calvignac-Spencer, S., Cobbold, C. A., Dawson, T. P., Emerson, B. C., Ferrier, S., Gilbert, M. T. P., Herold, M., Jones, L., Leendertz, F. H., Matthews, L., Millington, J. D. A., Olson, J. R., Ovaskainen, O., Raffaelli, D., Reeve, R., Rödel, M.-O., Rodgers, T. W., Snape, S., Visseren-Hamakers, I., Vogler, A. P., White, P. C. L., Wooster, M. J., and Yu, D. W. (2017). Connecting Earth observation to high-throughput biodiversity data. *Nature Ecology & Evolution*, 1(7):0176.

Carter, S. K., Fleishman, E., Leinwand, I. I. F., Flather, C. H., Carr, N. B., Fogarty, F. A., Leu, M., Noon, B. R., Wohlfeil, M. E., and Wood, D. J. A. (2019). Quantifying Ecological Integrity of Terrestrial Systems to Inform Management of Multiple-Use Public Lands in the United States. *Environmental Management*, 64(1):1–19.

Cavender-Bares, J., Schneider, F. D., Santos, M. J., Armstrong, A., Carnaval, A., Dahlin, K. M., Fatoyinbo, L., Hurtt, G. C., Schimel, D., Townsend, P. A., Ustin, S. L., Wang, Z., and Wilson, A. M. (2022).

Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nature Ecology & Evolution*, 6(5):506–519.

Chiaverini, L., Macdonald, D. W., Bothwell, H. M., Hearn, A. J., Cheyne, S. M., Haidir, I., Hunter, L. T. B., Kaszta, Z., Macdonald, E. A., Ross, J., and Cushman, S. A. (2022). Multi-scale, multivariate community models improve designation of biodiversity hotspots in the Sunda Islands. *Animal Conservation*, page acv.12771.

Christin, S., Hervet, E., and Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644.

Chua, P. Y., Bourlat, S. J., Ferguson, C., Korlevic, P., Zhao, L., Ekrem, T., Meier, R., and Lawniczak, M. K. (2023). Future of dna-based insect monitoring. *Trends in Genetics*, 39(7):531–544.

Davis, R. J., Ohmann, J. L., Kennedy, R. E., Cohen, W. B., Gregory, M. J., Yang, Z., Roberts, H. M., Gray, A. N., and Spies, T. A. (2015). Northwest Forest Plan–the first 20 years (1994-2013): status and trends of late-successional and old-growth forests. Technical Report PNW-GTR-911, U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, OR.

Diana, A., Matechou, E., Griffin, J., Yu, D. W., Luo, M., Tosa, M., Bush, A., and Griffiths, R. (2022). eDNAPlus: A unifying modelling framework for dna-based biodiversity monitoring. (arXiv:2211.12213). arXiv:2211.12213 [stat].

Dietz, T., Ostrom, E., and Stern, P. C. (2003). The struggle to govern the commons. *Science*, 302:1907–1912.

Elbrecht, V., Braukmann, T. W., Ivanova, N. V., Prosser, S. W., Hajibabaei, M., Wright, M., Zakharov, E. V., Hebert, P. D., and Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, 7:e7745.

Frankham, R. (2010). Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation*, 143(9):1919–1927.

Greenfield, P., Tran-Dinh, N., and Midgley, D. (2019). Kelpie: generating full-length 'amplicons' from whole-metagenome datasets. *PeerJ*, 6:e6174.

Guimarães Sales, N., McKenzie, M. B., Drake, J., Harper, L. R., Browett, S. S., Coscia, I., Wangensteen, O. S., Baillie, C., Bryce, E., Dawson, D. A., Ochu, E., Hänfling, B., Handley, L. L., Mariani, S., Lambin, X., Sutherland, C., and McDevitt, A. D. (2019). Fishing for mammals: landscape-level monitoring of terrestrial and semi-aquatic communities using eDNA from lotic ecosystems. preprint, Ecology.

Hamilton, H., Smyth, R. L., Young, B. E., Howard, T. G., Tracey, C., Breyer, S., Cameron, D. R., Chazal, A., Conley, A. K., Frye, C., and Schloss, C. (2022). Increasing taxonomic diversity and spatial resolution clarifies opportunities for protecting US imperiled species. *Ecological Applications*, 32(3).

Hartig, F., Abrego, N., Bush, A., Chase, J. M., Guillera-Arroita, G., Leibold, M. A., Ovaskainen, O., Pellissier, L., Pichler, M., Poggiato, G., Pollock, L., Si-Moussi, S., Thuiller, W., Viana, D. S., Warton, D., Zurell, D., and Yu, D. W. (2023). Novel community data – properties and prospects.

He, K. S., Bradley, B. A., Cord, A. F., Rocchini, D., Tuanmu, M., Schmidtlein, S., Turner, W., Wegmann, M., and Pettorelli, N. (2015). Will remote sensing shape the next generation of species distribution models? *Remote Sensing in Ecology and Conservation*, 1(1):4–18.

Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through dna barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512):313–321.

Hobbs, R. J., Cole, D. N., Yung, L., Zavaleta, E. S., Aplet, G. H., Chapin, F. S., Landres, P. B., Parsons, D. J., Stephenson, N. L., White, P. S., Graber, D. M., Higgs, E. S., Millar, C. I., Randall, J. M., Tonnessen, K. A., and Woodley, S. (2010). Guiding concepts for park and wilderness stewardship in an era of global environmental change. *Frontiers in Ecology and the Environment*, 8(9):483–490.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P. M., Woodcock, P., Edwards, F. A., Larsen, T. H., Hsu, W. W., Benedick, S., Hamer, K. C., Wilcove, D. S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B. C., and Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16(10):1245–1257.

Ji, Y., Baker, C. C. M., Popescu, V. D., Wang, J., Wu, C., Wang, Z., Li, Y., Wang, L., Hua, C., Yang, Z., Yang, C., Xu, C. C. Y., Diana, A., Wen, Q., Pierce, N. E., and Yu, D. W. (2022). Measuring protected-area effectiveness using vertebrate distributions from leech iDNA. *Nature Communications*, 13(1):1555.

Ji, Y., Huotari, T., Roslin, T., Schmidt, N. M., Wang, J., Yu, D. W., and Ovaskainen, O. (2020). SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Molecular Ecology Resources*, 20(1):256–267.

Kane, V. R., McGaughey, R. J., Bakker, J. D., Gersonde, R. F., Lutz, J. A., and Franklin, J. F. (2010). Comparisons between field- and LiDAR-based measures of stand structural complexity. *Canadian Journal of Forest Research*, 40(4):761–773.

Krijthe, J. H. (2015). *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*. R package version 0.15.

Kukkala, A. S. and Moilanen, A. (2013). Core concepts of spatial prioritisation in systematic conservation planning. *Biological Reviews*, 88(2):443–464.

Kwok, R. (2018). Ecology's remote-sensing revolution. *Nature*, 556(7699):137–138.

Leempoel, K., Hebert, T., and Hadly, E. A. (2019). A comparison of eDNA to camera trapping for assessment of terrestrial mammal diversity. preprint, Ecology.

Leitão, P. J. and Santos, M. J. (2019). Improving Models of Species Ecological Niches: A Remote Sensing Overview. *Frontiers in Ecology and Evolution*, 7:9.

Lewinsohn, T. M. and Roslin, T. (2008). Four ways towards tropical herbivore megadiversity. *Ecology Letters*, 11(4):398–416.

Lin, M., Simons, A. L., Harrigan, R. J., Curd, E. E., Schneider, F. D., Ruiz-Ramos, D. V., Gold, Z., Osborne, M. G., Shirazi, S., Schweizer, T. M., Moore, T. N., Fox, E. A., Turba, R., Garcia-Vedrenne, A. E., Helman, S. K., Rutledge, K., Mejia, M. P., Marwayana, O., Munguia Ramos, M. N., Wetzer, R., Pentcheff, N. D., McTavish, E. J., Dawson, M. N., Shapiro, B., Wayne, R. K., and Meyer, R. S. (2021). Landscape analyses using edna metabarcoding and earth observation predict community biodiversity in california. *Ecological Applications*, 31(6):e02379.

Lock, M., van Duren, I., Skidmore, A. K., and Saintilan, N. (2022). Harmonizing Forest Conservation Policies with Essential Biodiversity Variables Incorporating Remote Sensing and Environmental DNA Technologies. *Forests*, 13(3):445.

Loomis, J. (2002). *Integrated Public Lands Management: Principles and Applications to National Forests, Parks, Wildlife Refuges, and BLM Lands*. Columbia University Press.

Luo, M., Ji, Y., Warton, D., and Yu, D. W. (2023). Extracting abundance information from DNA-based data. *Molecular Ecology Resources*, 23(1):174–189.

Lyet, A., Pellissier, L., Valentini, A., Dejean, T., Hehmeyer, A., and Naidoo, R. (2021). eDNA sampled from stream networks correlates with camera trap detection rates of terrestrial mammals. *Scientific Reports*, 11(1):11362.

Massey, A. L., Bronzoni, R. V. d. M., Silva, D. J. F., Allen, J. M., Lázari, P. R., Santos-Filho, M., Canale, G. R., Bernardo, C. S. S., Peres, C. A., and Levi, T. (2022). Invertebrates for vertebrate biodiversity monitoring: Comparisons using three insect taxa as iDNA samplers. *Molecular Ecology Resources*, 22(3):962–977.

Maxwell, S. L., Cazalis, V., Dudley, N., Hoffmann, M., Rodrigues, A. S. L., Stolton, S., Visconti, P., Woodley, S., Kingston, N., Lewis, E., Maron, M., Strassburg, B. B. N., Wenger, A., Jonas, H. D., Venter, O., and Watson, J. E. M. (2020). Area-based conservation in the twenty-first century. *Nature*, 586(7828):217–227.

Mayer, M. (2021). *flashlight: Shed Light on Black Box Machine Learning Models*. R package version 0.8.0.

Metcalfe, P., Beven, K., and Freer, J. (2018). *dynatopmodel: Implementation of the Dynamic TOPMODEL Hydrological Model*.

Müller, J., Bae, S., Röder, J., Chao, A., and Didham, R. K. (2014). Airborne LiDAR reveals context dependence in the effects of canopy architecture on arthropod diversity. *Forest Ecology and Management*, 312:129–137.

Müller, J. and Brandl, R. (2009). Assessing biodiversity by remote sensing in mountainous terrain: the potential of LiDAR to predict forest beetle assemblages. *Journal of Applied Ecology*, 46(4):897–905.

Müller, J., Moning, C., Bässler, C., Heurich, M., and Brandl, R. (2009). Using airborne laser scanning to model potential abundance and assemblages of forest passerines. *Basic and Applied Ecology*, 10(7):671–681.

Ovaskainen, O. and Abrego, N. (2020). *Joint Species Distribution Modelling: With Applications in R*. Cambridge University Press, 1 edition.

Pawlowski, J., Apothéloz-Perret-Gentil, L., and Altermatt, F. (2020). Environmental DNA: What's behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Molecular Ecology*, 29(22):4258–4264.

Pettorelli, N., Schulte to Bühne, H., Tulloch, A., Dubois, G., Macinnis-Ng, C., Queirós, A. M., Keith, D. A., Wegmann, M., Schrodt, F., Stellmes, M., Sonnenschein, R., Geller, G. N., Roy, S., Somers, B., Murray, N., Bland, L., Geijzendorffer, I., Kerr, J. T., Broszeit, S., Leitão, P. J., Duncan, C., El Serafy, G., He, K. S., Blanchard, J. L., Lucas, R., Mairota, P., Webb, T. J., and Nicholson, E. (2018). Satellite remote sensing of ecosystem functions: opportunities, challenges and way forward. *Remote Sensing in Ecology and Conservation*, 4(2):71–93.

Pichler, M. and Hartig, F. (2021). A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods in Ecology and Evolution*, 12(11):2159–2173.

Pichler, M. and Hartig, F. (2023). Machine learning and deep learning—a review for ecologists. *Methods in Ecology and Evolution*, 14(4):994–1016.

Pollock, L. J., O'Connor, L. M., Mokany, K., Rosauer, D. F., Talluto, M. V., and Thuiller, W. (2020). Protecting Biodiversity (in All Its Complexity): New Models and Methods. *Trends in Ecology & Evolution*, 35(12):1119–1128.

Prather, C. M., Pelini, S. L., Laws, A., Rivest, E., Woltz, M., Bloch, C. P., Del Toro, I., Ho, C.-K., Kominoski, J., Newbold, T. A. S., Parsons, S., and Joern, A. (2013). Invertebrates, ecosystem services and climate change: Invertebrates, ecosystems and climate change. *Biological Reviews*, 88(2):327–348.

Ratnasingham, S. (2019). mbrave: The multiplex barcode research and visualization environment. *Biodiversity Information Science and Standards*, 3:e37986.

Rhodes, M. W., Bennie, J. J., Spalding, A., ffrench-Constant, R. H., and Maclean, I. M. D. (2022). Recent advances in the remote sensing of insects. *Biological Reviews*, 97(1):343–360.

Rodgers, T. W., Xu, C. C. Y., Giacalone, J., Kapheim, K. M., Saltonstall, K., Vargas, M., Yu, D. W., Somervuo, P., McMillan, W. O., and Jansen, P. A. (2017). Carrion fly-derived dna metabarcoding is an effective tool for mammal surveys: Evidence from a known tropical mammal community. *Molecular Ecology Resources*, 17(6):e133–e145.

Ruppert, K. M., Kline, R. J., and Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, 17:e00547.

Speaker, T., O'Donnell, S., Wittemyer, G., Bruyere, B., Loucks, C., Dancer, A., Carter, M., Fegraus, E., Palmer, J., Warren, E., and Solomon, J. (2022). A global community-sourced assessment of the state of conservation technology. *Conservation Biology*, 36(3).

Srivathsan, A., Lee, L., Katoh, K., Hartop, E., Kutty, S. N., Wong, J., Yeo, D., and Meier, R. (2021). Ontbarcoder and minion barcodes aid biodiversity discovery and identification by everyone, for everyone. *BMC Biology*, 19(1):217.

Thomsen, P. F. and Sigsgaard, E. E. (2019). Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. *Ecology and Evolution*, 9(4):1665–1679.

Tilker, A., Abrams, J. F., Nguyen, A., Hörig, L., Axtner, J., Louvrier, J., Rawson, B. M., Nguyen, H. A. Q., Guegan, F., Nguyen, T. V., Le, M., Sollmann, R., and Wilting, A. (2020). Identifying conservation priorities in a defaunated tropical biodiversity hotspot. *Diversity and Distributions*, 26(4):426–440.

Tosa, M. I., Dziedzic, E. H., Appel, C. L., Urbina, J., Massey, A., Ruprecht, J., Eriksson, C. E., Dolliver, J. E., Lesmeister, D. B., Betts, M. G., Peres, C. A., and Levi, T. (2021). The Rapid Rise of Next-Generation Natural History. *Frontiers in Ecology and Evolution*, 9:698131.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., and Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, 7(1):9132.

van der Maaten, L. and Hinton, G. (2008). Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

van Klink, R., August, T., Bas, Y., Bodesheim, P., Bonn, A., Fossøy, F., Høye, T. T., Jongejans, E., Menz, M. H. M., Miraldo, A., Roslin, T., Roy, H. E., Ruczyński, I., Schigel, D., Schäffler, L., Sheard, J. K., Svenningsen, C., Tschan, G. F., Wäldchen, J., Zizka, V. M. A., Åström, J., and Bowler, D. E. (2022). Emerging technologies revolutionise insect ecology and monitoring. *Trends in Ecology & Evolution*.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., and Hui, F. K. (2015). So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, 30(12):766–779.

Westgate, M. J., Barton, P. S., Lane, P. W., and Lindenmayer, D. B. (2014). Global meta-analysis reveals low consistency of biodiversity congruence relationships. *Nature Communications*, 5(1):3899.

Weston, P. (2021). New biodiversity algorithm 'will blight range of natural habitats in England'. *The*

Guardian.

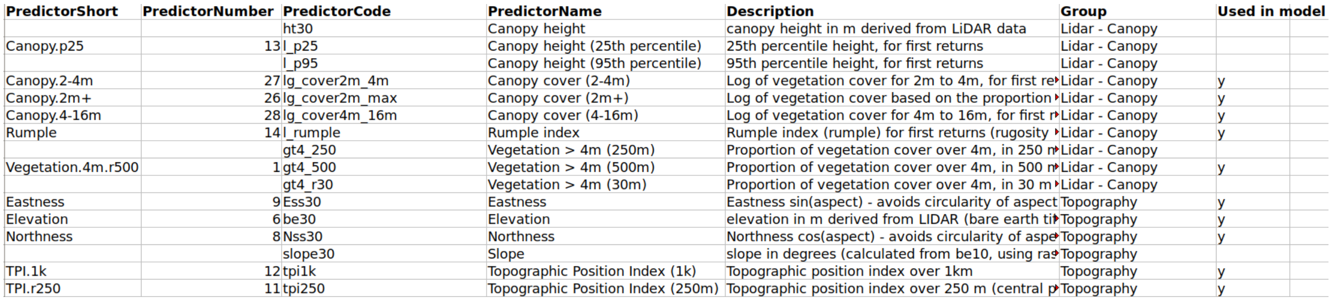
Wilson, M. F. J., O'Connell, B., Brown, C., Guinan, J. C., and Grehan, A. J. (2007). Multiscale Terrain Analysis of Multibeam Bathymetry Data for Habitat Mapping on the Continental Slope. *Marine Geodesy*, 30(1-2):3–35.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition.

Zhang, K., Lin, S., Ji, Y., Yang, C., Wang, X., Yang, C., Wang, H., Jiang, H., Harrison, R. D., and Yu, D. W. (2016). Plant diversity accurately predicts insect diversity in two tropical landscapes. *Molecular Ecology*, 25(17):4407–4419.

Supplementary Figures for the Article 'Combining environmental DNA and remote sensing for efficient, fine-scale mapping of arthropod biodiversity'

Yuanheng Li, Christian Devenish, Marie I. Tosa, Mingjie Luo, David M. Bell, Damon B. Lesmeister, Paul Greenfield, Maximilian Pichler, Taal Levi, and Douglas W. Yu



PredictorShort	PredictorNumber	PredictorCode	PredictorName	Description	Group	Used in model
		ht30	Canopy height	canopy height in m derived from LiDAR data	Lidar - Canopy	
Canopy.p25	13	l_p25	Canopy height (25th percentile)	25th percentile height, for first returns	Lidar - Canopy	
		l_p95	Canopy height (95th percentile)	95th percentile height, for first returns	Lidar - Canopy	
Canopy.2-4m	27	lg_cover2m_4m	Canopy cover (2-4m)	Log of vegetation cover for 2m to 4m, for first re	Lidar - Canopy	y
Canopy.2m+	26	lg_cover2m_max	Canopy cover (2m+)	Log of vegetation cover based on the proportion	Lidar - Canopy	y
Canopy.4-16m	28	lg_cover4m_16m	Canopy cover (4-16m)	Log of vegetation cover for 4m to 16m, for first r	Lidar - Canopy	y
Rumple	14	l_rumple	Rumple index	Rumple index (rumple) for first returns (rugosity	Lidar - Canopy	y
		gt4_250	Vegetation > 4m (250m)	Proportion of vegetation cover over 4m, in 250 m	Lidar - Canopy	
Vegetation.4m.r500	1	gt4_500	Vegetation > 4m (500m)	Proportion of vegetation cover over 4m, in 500 m	Lidar - Canopy	y
		gt4_r30	Vegetation > 4m (30m)	Proportion of vegetation cover over 4m, in 30 m	Lidar - Canopy	
Eastness	9	Ess30	Eastness	Eastness sin(aspect) - avoids circularity of aspect	Topography	y
Elevation	6	be30	Elevation	elevation in m derived from LIDAR (bare earth ti	Topography	y
Northness	8	Nss30	Northness	Northness cos(aspect) - avoids circularity of aspe	Topography	y
		slope30	Slope	slope in degrees (calculated from be10, using ras	Topography	
TPI.1k	12	tpi1k	Topographic Position Index (1k)	Topographic position index over 1km	Topography	y
TPI.r250	11	tpi250	Topographic Position Index (250m)	Topographic position index over 250 m (central p	Topography	y

Table 1S: All candidate predictors for jSDM model. Predictors are grouped by origin: Lidar, Landsat, H.J. Andrews Experimental Forest GIS data; 29 predictors were included in the model, chosen by Variance Inflation Factor (VIF) < 8, as well as the categorical predictor of inside or outside the boundaries of H.J. Andrews Experimental Forest. Elevation was forced to be included regardless of VIF value. The full table is in `https://github.com/chnpenny/HJA_analyses_Kelpie_clean/blob/main/05_supplement/GIS/Table_1S.xlsx`

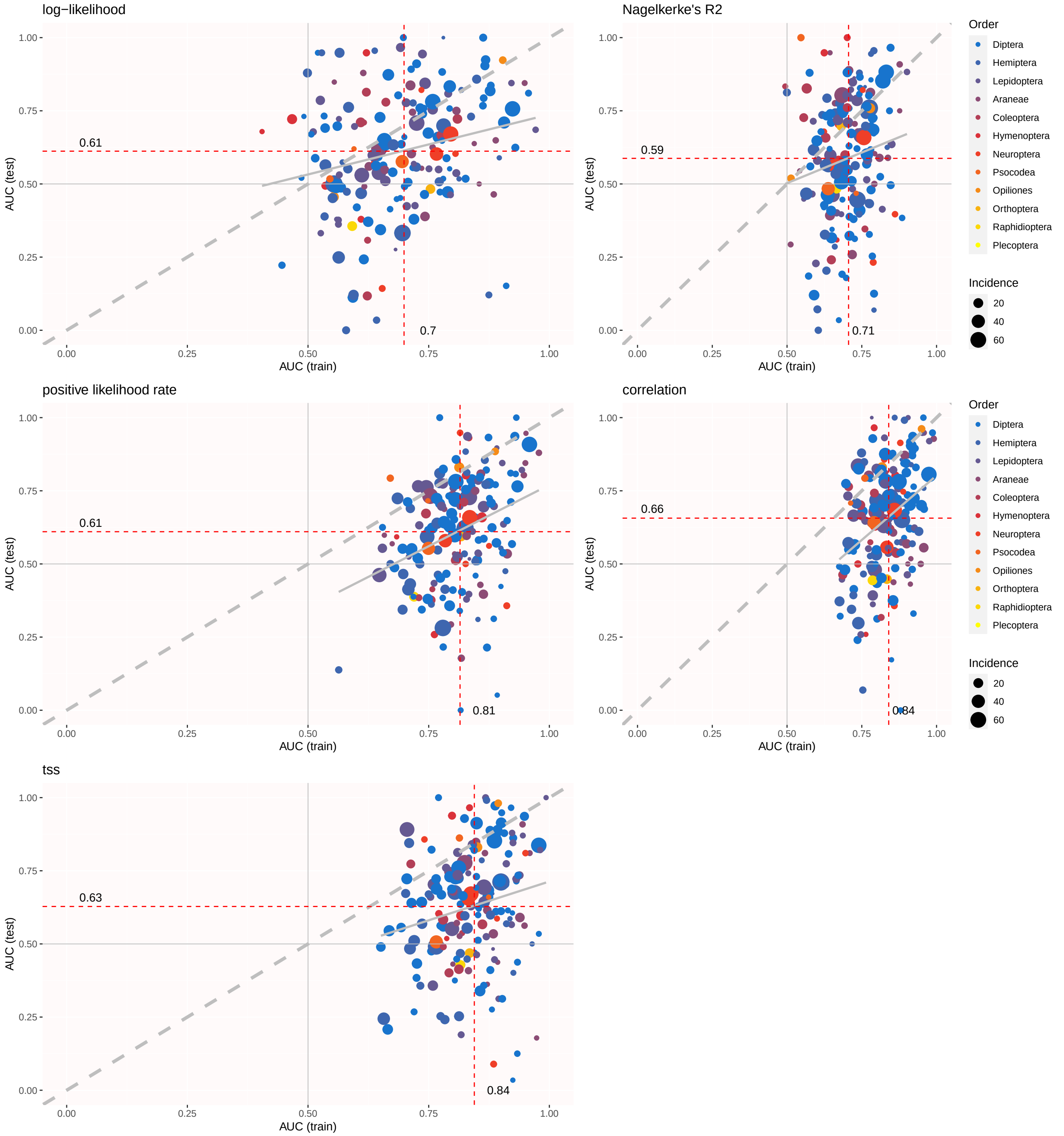


Figure 1S: Explanatory AUC vs predictive AUC for best sjSDM models tuned according to log-likelihood, Nagelkerke's R^2, positive likelihood rate, correlation and TSS(true skill statistic). Each point is one OTU. Color indicates taxonomic class (order), and point size indicates incidence (number of Malaise traps in which the OTU was detected). The dashed gray line is the 1 : 1 line, and the solid gray line is a fitted linear regression.

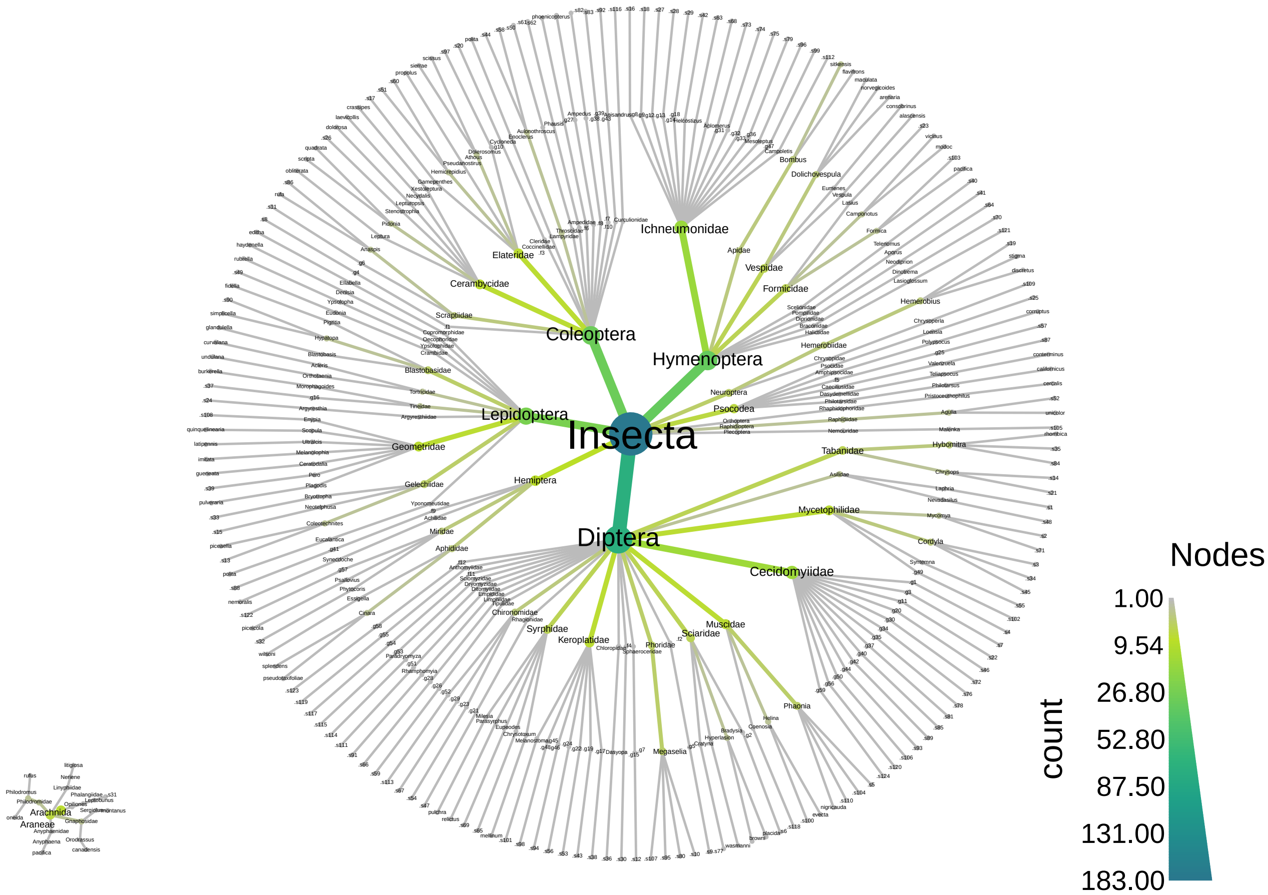


Figure 2S: Detailed taxonomic distribution of 190 Operational Taxonomic Units (OTUs) over two heat trees, the Insecta and the Arachnida. Node size and color are scaled to the number of OTUs in that node. Missing taxonomic information of species are indicated by the combination of a point, f, g or s, representing family, genus or species, respectively, and a number, e.g. '.f15'.

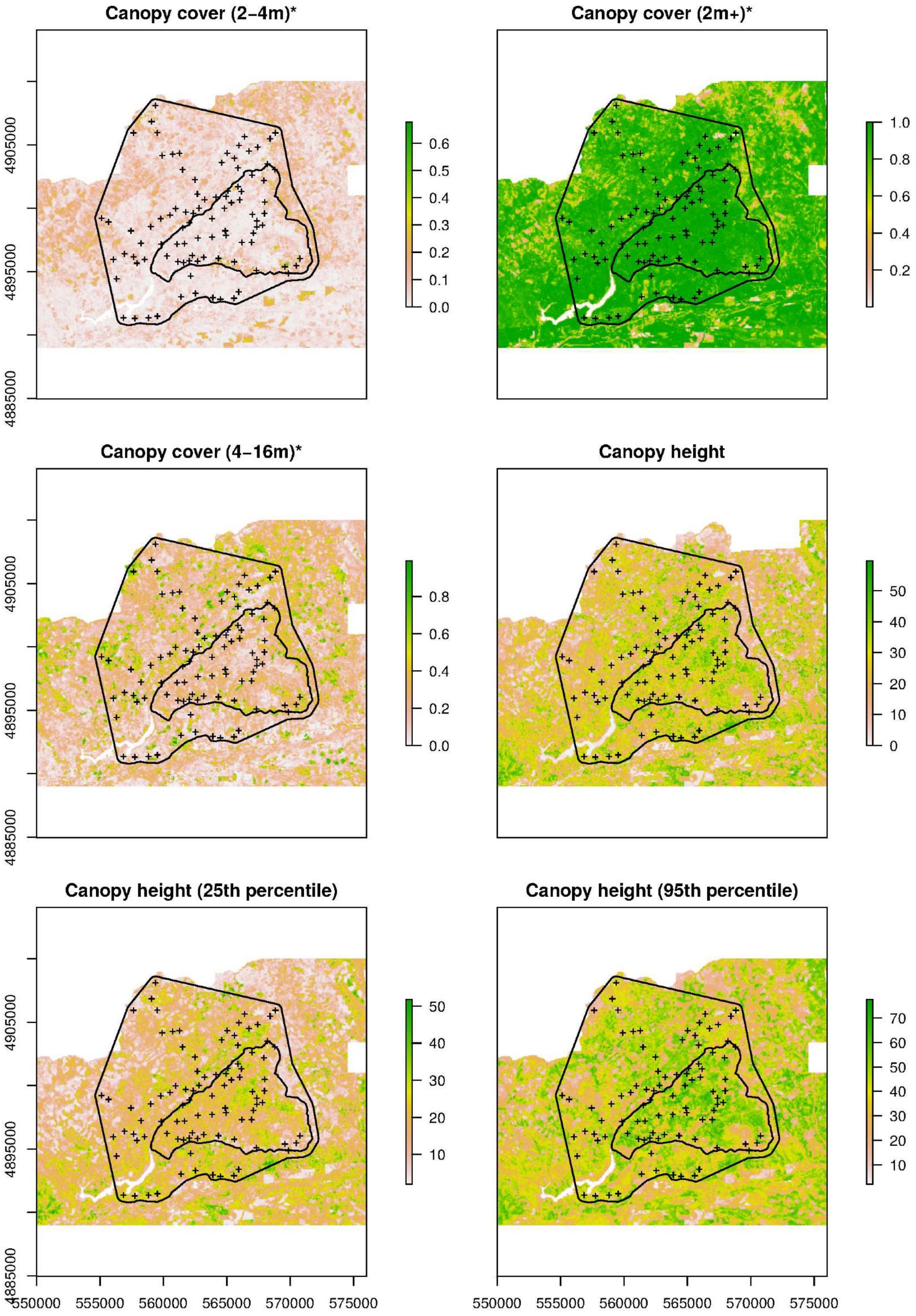


Figure 3S: All candidate covariates. Sample locations are marked by the plus sign, inner black outline shows H.J. Andrews Experimental Forest boundary and outer black outline shows extent of prediction area, Covariates used in model are marked with an asterisk. See Table S-covariates for covariate descriptions. The full figures are in `https://github.com/chnpenny/HJA_analyses_Kelpie_clean/blob/main/05_supplement/Plots/Figure_3S-full.pdf`.

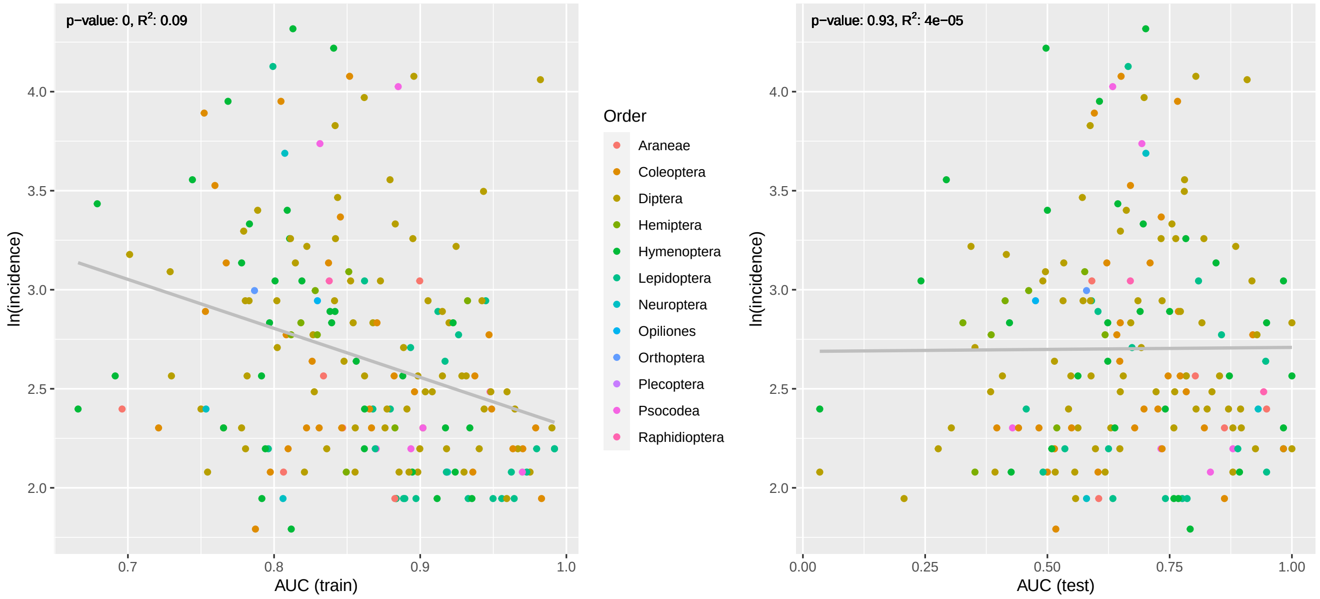


Figure 4S: Explanatory (training) and predictive (test) AUCs of all OTUs by incidence. Colors correspond with the order of OTUs. OTUs that are detected less (low incidence) show larger variance in the AUC values. The p-value and R^2 of the linear regressions are shown on the top of the plots. To be noticed, incidences of OTUs are log-transformed.

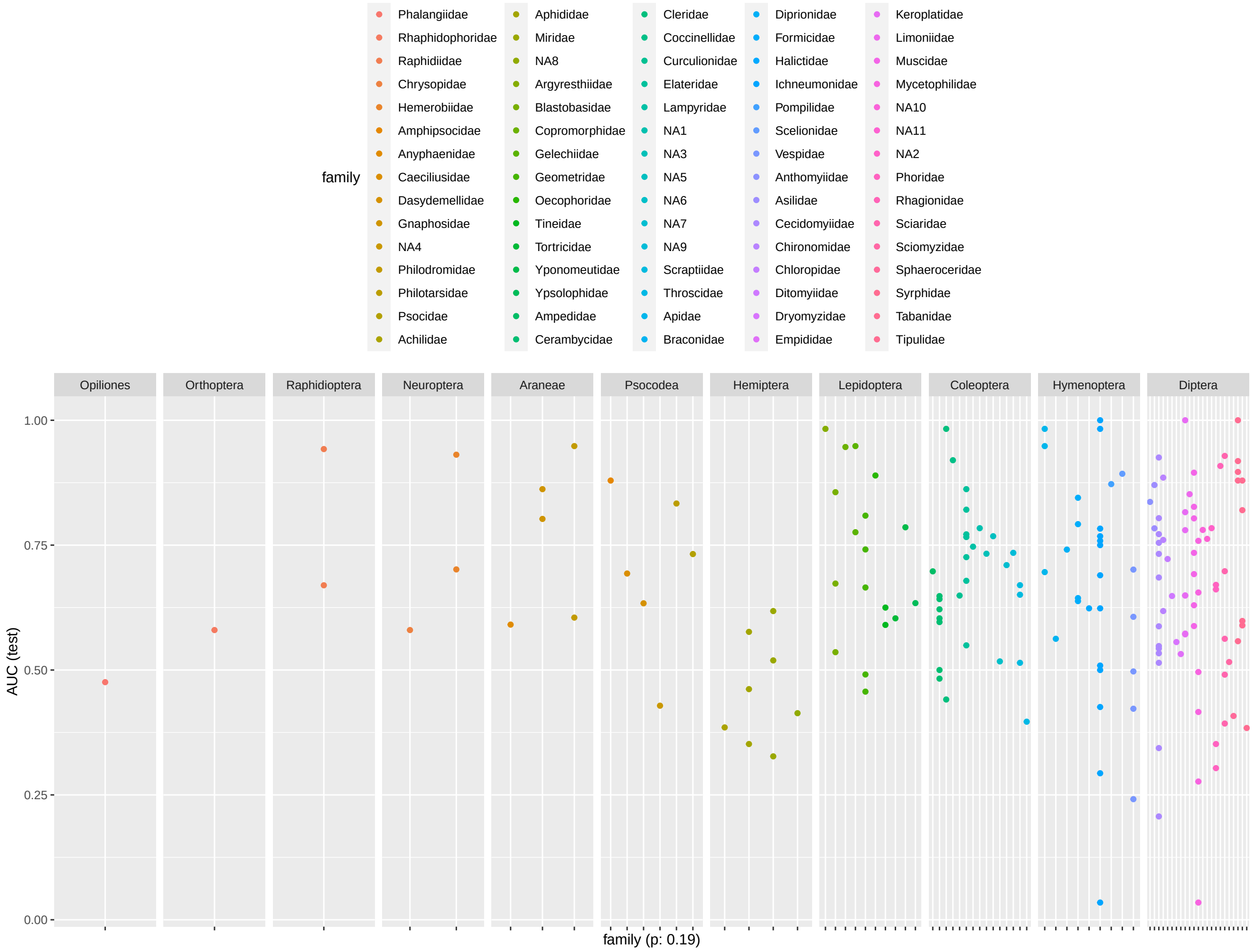


Figure 5S: Predictive AUCs of all OTUs by taxonomic family. Colors correspond with the family information and they are arranged according to the order information. A linear regression shows that there is no significant effect of family on the predictive AUCs (p-value 0.19 for this regression).

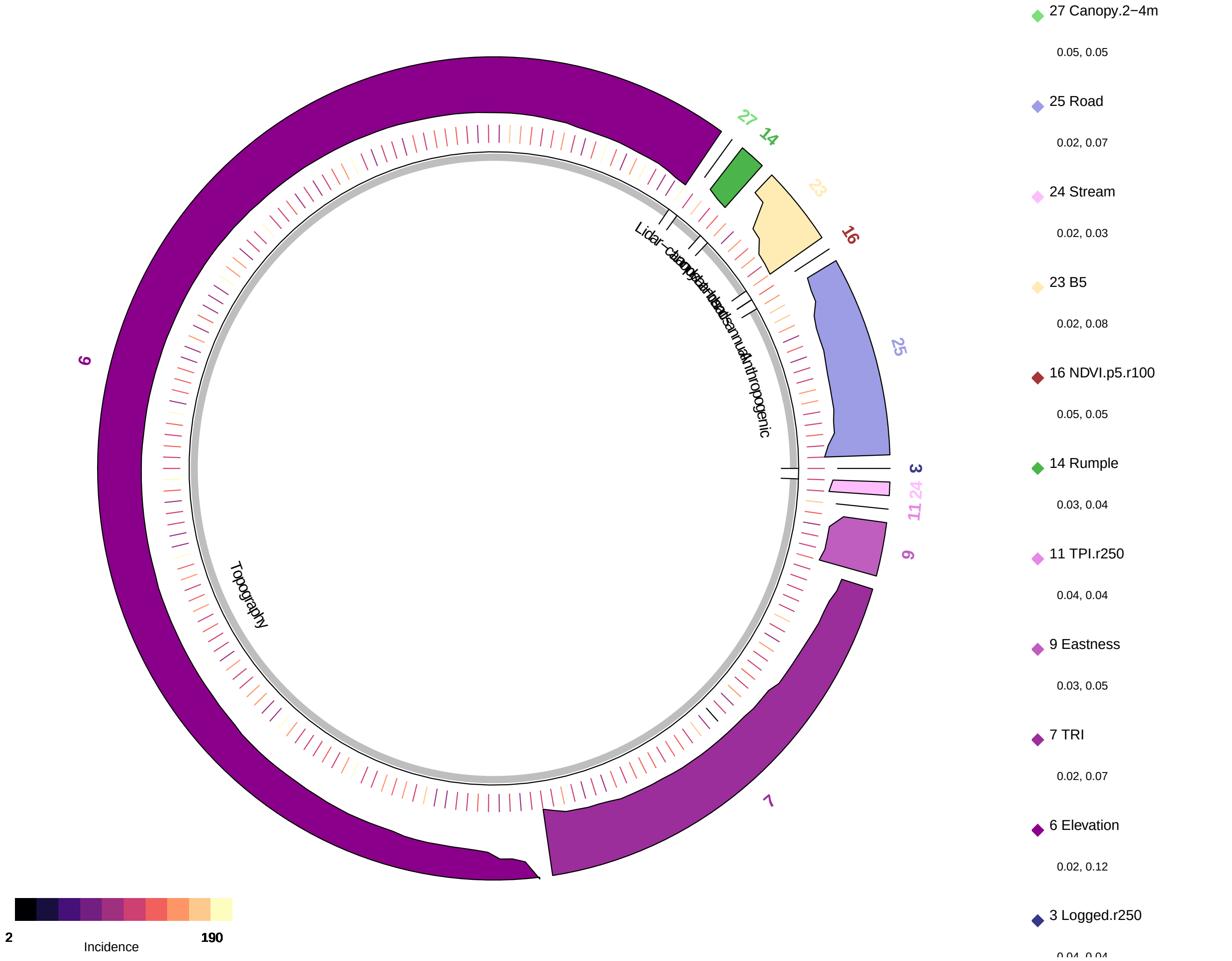


Figure 6S: Variable importance for interaction effects. The importance of environmental covariates for each OTU with regard to their interaction effects, excluding spatial location variables. Tick marks indicate OTU incidence, color bands indicate individual covariates, and gray bands indicate covariate groupings (Table 1S). Elevation (variable 6) and TRI (variable 7) are the most important variables. The heights of the colour bars are scaled to the Friedman's H statistic for overall interaction strength for that OTU. The ranges of overall interaction strength for each variable are shown in the legend on the right.

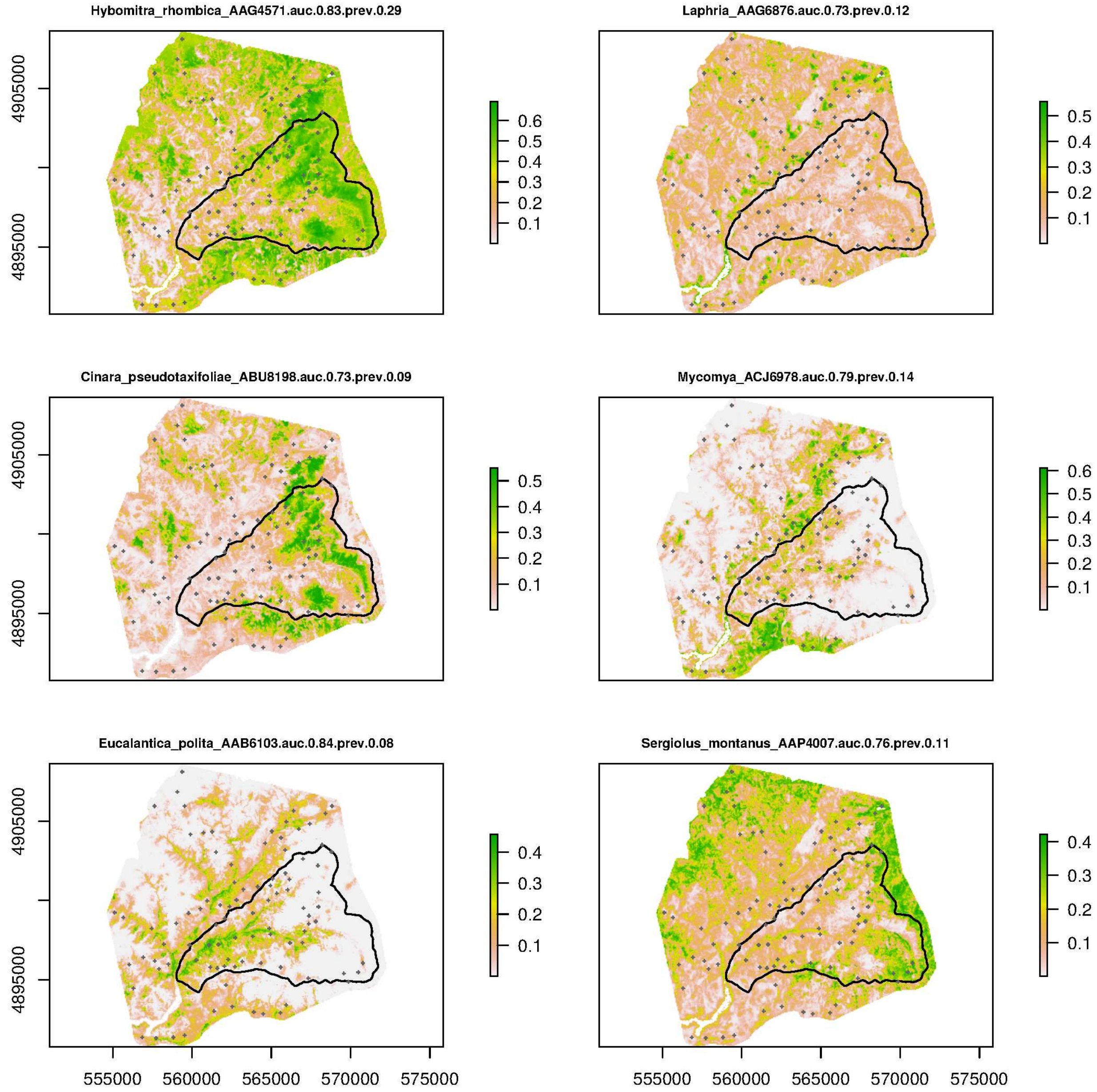

Hybomitra_rhombica_AAG4571.auc.0.83.prev.0.29
4905000
4895000
0.6
0.5
0.4
0.3
0.2
0.1
Laphria_AAG6876.auc.0.73.prev.0.12
0.5
0.4
0.3
0.2
0.1
Cinara_pseudotaxifoliae_ABU8198.auc.0.73.prev.0.09
4905000
4895000
0.5
0.4
0.3
0.2
0.1
Mycomya_ACJ6978.auc.0.79.prev.0.14
0.6
0.5
0.4
0.3
0.2
0.1
Eucalantica_polita_AAB6103.auc.0.84.prev.0.08
4905000
4895000
0.4
0.3
0.2
0.1
555000
560000
565000
570000
575000
Sergiolus_montanus_AAP4007.auc.0.76.prev.0.11
0.4
0.3
0.2
0.1
555000
560000
565000
570000
575000

Figure 7S: Individual, interpolated species distributions. The full figure is in `https://github.com/chnpenny/HJA_analyses_Kelpie_clean/blob/main/05_supplement/Plots/Figure_7S-full.pdf`

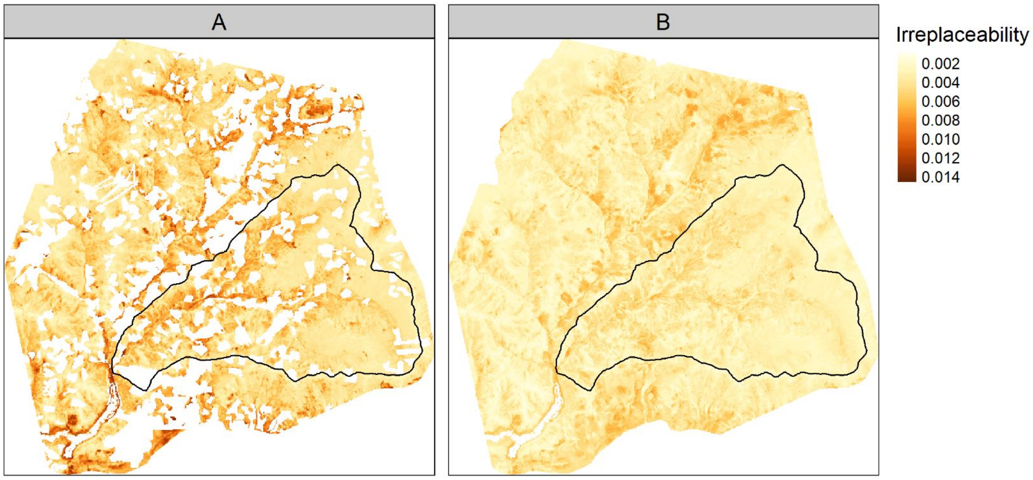


Figure 8S: Site-irreplaceability values plotted across the study area, showing HJA Experimental Forest boundaries (black line). A. With plantations masked out. B. With plantations present. Note the higher irreplaceability values in plantations, given that species mainly restricted to plantations are rarer across our study area than those in old growth forests.

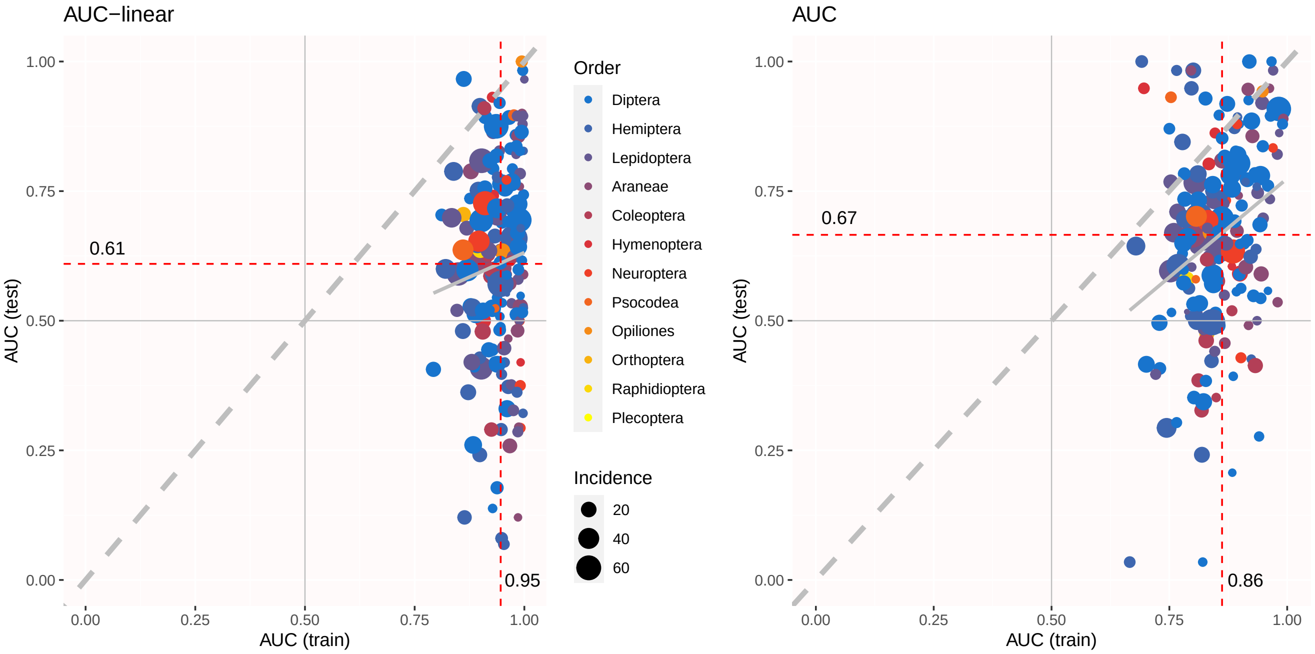


Figure 9S: Explanatory and predictive AUCs of the tuned sjSDM model applying linear fitting on the environmental part (left panel) to the same model applying DNN fitting (right panel). The explanatory power (x axis, AUC (train)) is higher but the predictive power (y axis, AUC (test)) is lower in the linear model, relative to the DNN model.

Supplementary Information for the Article 'Combining environmental DNA and remote sensing for efficient, fine-scale mapping of arthropod biodiversity': Materials and Methods

Yuanheng Li, Christian Devenish, Marie I. Tosa, Mingjie Luo, David M. Bell, Damon B. Lesmeister, Paul Greenfield, Maximilian Pichler, Taal Levi, and Douglas W. Yu

Dietz's five elements and the creation of a biodiversity offset market

A rare example of all five elements working together to achieve biodiversity conservation is the UK District Licensing offset market for the great crested newt (*Triturus cristatus*). Until recently, builders had been required to survey for the newt when their plans might affect ponds, and to respond to newt detections by paying for mitigation measures. Traditional surveys required at least four visits per pond during the short breeding season. After Biggs et al. [2015] showed that a single environmental-DNA (eDNA) water survey per pond, analysed with probe-based quantitative PCR (qPCR), could detect the newt with equal sensitivity (i.e. eDNA information is *high-quality* and *granular*), the UK government authorised newt eDNA surveys, and a private laboratory market grew to *provide the infrastructure* for *timely* and *trustworthy* information, via response times of a few days and an annual proficiency test. The switch to eDNA increased survey efficiency, but still left in place the UK's reactive approach to newt conservation ('mitigate after impact'). Mitigation measures, such as translocation, can delay building by over a year. In 2018, the UK government took further advantage of eDNA's detection efficiency by implementing an *institutional redesign* with the District Licensing scheme, where hundreds of ponds across one or more local planning authorities are first systematically surveyed with eDNA [Natural England, 2019]. The data are used to fit a species distribution model, which is converted to an *understandable* map of discrete risk zones for the newt. Builders can now meet their legal obligations at any time by paying for a license, the cost of which depends on their site's size, risk-zone level, and number of affected ponds, eliminating delay. The licence fees fund the proactive creation and long-term management of compensation habitat, including four new ponds per affected pond. Compensation habitat is directed toward Strategic Opportunity Areas, which reflect planning-authority building aspirations (*political bargaining*), and *enforcement* is through the same processes that apply to all planning permissions.

Materials and Methods

Model Inputs

Field data collection

We collected 121 Malaise-trap samples of arthropods at 89 sampling sites in and around the H.J. Andrews Experimental Forest and Long-Term Ecological Research site (HJA), Oregon, USA in July 2018. Sites were stratified (as best as possible while yielding to logistical constraints) based on elevation and time since disturbance. Sites were also stratified between inside and outside the HJA to capture landscape-scale differences between a long-term ecological research site where no logging has occurred since 1989 and neighboring sites within a landscape context with continued active management. Each trap was left to collect for seven days, and samples were transferred to fresh 100% ethanol to store at room temperature until extraction. In 32 of the sites, two Malaise traps were set 40 m apart, and in the other 57, only one trap was set (Figure 1M A). In August 2018, we repeated the sampling and processed all 242 samples together, but we have analyzed only the July samples for this study.

Wet-lab pipeline and bioinformatics

We follow the SPIKEPIPE protocol from Ji et al. [2020], where we map paired-end reads from Illumina shotgun-sequenced samples to a reference dataset of DNA barcode sequences. In shotgun sequencing, the total DNA of each sample is sequenced (the term shotgun refers to the random subset of the total DNA that gets sequenced), and the output 'reads' are 'mapped' (matched) to a reference set of barcodes. This approach relies on the enormous data output of Illumina sequencers, since only $\sim 1/4000$ reads is from a DNA barcode, as opposed to the rest of the genome.

A major benefit of the SPIKEPIPE method is reduced workload since all that is needed is to extract DNA from each sample before sending to a sequencing center. The main disadvantage is that species present at low overall biomass are unlikely to be detected (although this is also a partial advantage in that any sample cross-contamination is also unlikely to be detected). However, low-biomass species are less likely to contribute meaningfully to species distribution modelling since the numbers of incidences for rare species are, by definition, low.

An important difference of this study from Ji et al. [2020] is that their study used a pre-existing reference set of DNA barcodes [Wirta et al., 2014], whereas we generate our reference set directly from the same shotgun-sequenced datasets, using the program Kelpie [Greenfield et al., 2019], which is an *in-silico* PCR program.

For this study, we only analyzed the July 2018 samples ($n = 121$), but the arthropod samples of both sessions were together extracted, sequenced, analyzed, and assigned to taxonomies.

DNA extraction and sequencing

DNA was non-destructively extracted by soaking the samples in 5X lysis buffer while shaking and incubating the samples at 56 °C for 60 h [for more details, see Ji et al., 2020]. To the lysis buffers, we added a DNA spike-in standard of two beetle species in a 9 : 1 ratio. We shotgun-sequenced all 242 samples (PE 150, 350 bp insert size) to a mean depth of 29.0 million read pairs (range 21-47) on an Illumina NovaSeq 6000 at Novogene (Beijing, China). We used `TrimGalore 0.4.5` (`https://www.bioinformatics.babraham.ac.uk/projects/trim_galore`, accessed 10 Sep 2021) to remove residual adapters (`--paired --length 100 -trim-n`).

Creating a barcode reference database using *Kelpie in-silico* PCR

In physical PCR, two specially designed DNA sequences known as PCR primers are used to amplify (make many copies of) a target sequence, which, here, is the portion of the mitochondrial cytochrome oxidase subunit I (COI) gene that is widely used as the taxonomically informative 'DNA barcode'. If we had tried to use physical PCR to construct a reference library of DNA barcodes from the Malaise trap sample set, we would have needed to individually separate, sort, identify, extract, and PCR many hundreds of specimens.

Instead, we used a recently available shortcut known as 'in-silico PCR', using a software package called `Kelpie` [Greenfield et al., 2019]. Using the shotgun-sequence read files from the Malaise-trap samples, `Kelpie` carries out a computer search for reads that match the two ends of the target DNA barcode and then searches for overlapping reads, ultimately assembling DNA barcode sequences from the shotgun datasets. In our case, we use the BF3+BR2 primers from Elbrecht et al. [2019], which bookend a 418-bp fragment of the COI DNA barcode. After running `Kelpie` on all individual and groups of Malaise trap samples, `Kelpie` assembled 5560 unique DNA-barcode sequences. some more abundant than others.

We first used `FilterReads` to reduce the shotgun datasets to reads that resemble COI sequences, using a reference kmer dataset `GenBank_24919_COI_C99_20.mer` (accessed 3 Aug 2021). This step is optional but greatly increases efficiency (`FilterReads -qt 30 +f GenBank_24919_COI_C99_20.mer 25pct input.fq`). We then used `Kelpie 2.0.11` (Greenfield et al. 2019) to carry out *in-silico* PCR on the filtered datasets. Binaries for both are at `https://github.com/PaulGreenfieldOz/WorkingDogs` (accessed 3 Aug 2021). Kelpie mimics PCR on shotgun datasets by finding reads that include the forward primer sequence and step-by-step overlapping reads until a read matching the reverse primer is found (`Kelpie -f CCHGAYATRGCHTTYCCHCG -r TCDGGRTGNCCRAARAAYCA -primers -filtered -min 400 -max 500`). The advantages are that it is trivial to switch primers, workload is reduced, there can be no PCR error or contamination, and the primer regions are returned.

The main disadvantage of *Kelpie* is that low-abundance species in a sample are usually not detected since every species requires enough reads in the dataset to complete the assembly from the forward to the reverse primer.

That said, low-biomass OTUs are unlikely to contribute much to modelling, as they are also likely to exhibit low prevalence (few detection events) in the dataset. Nonetheless, we still tried to retrieve as many OTUs as possible by running *Kelpie* individually on each of the 242 samples and also running on concatenated fastq files made up of sample clusters (each site and its five nearest neighbors). The logic for the two steps is that even rare species might be abundant somewhere. In our experience, it is not helpful to concatenate large numbers of sequence files because rare amplicons look like error variants when there also exists in the dataset a similar but abundant amplicon sequence. *Kelpie* removes such rare amplicons as part of its error correction procedure. We combined the *Kelpie* outputs, gave the sequences unique names, and dereplicated, resulting in 5560 unique sequences.

The variation represented by these 5560 unique sequences derives from multiple causes: true genetic differences among species, true genetic diversity within species, errors generated by the Illumina sequencer, and rare pseudogene sequences from mitochondrial DNA that got copied into the nuclear genome at various points in each species' past and been released from purifying selection. The latter are known as NUMTs (nuclear mitochondrial DNA).

We assigned taxonomies to all 5560 unique sequences on `https://www.gbif.org/tools/sequence-id` (accessed 3 Aug 2021), which provides three sequence-match classes (`'exact'`, `'close'`, and `'no'` match). For the exact match class, we retained the assignment to species, for the close match class, we retained the assigned genus and used NA for the species epithet, and for the weak match class, we retained the assigned order and used NA for lower ranks. We deleted all sequences that received a `'no match'` or were not assigned to Insecta or Arachnida, after which, we used `vsearch 2.15.0` to cluster the sequences into 1538 97%-similarity OTUs.

Although PCR error has been avoided, Kelpie amplicons unavoidably still include Illumina sequencer error, including homopolymers (incorrect nucleotide repeats), which induce frameshift mutations. However, because the amplicon is of a protein-coding gene, we aligned the OTU representative sequences by their inferred amino-sequences ('translation alignment'), using the invertebrate mitochondrial code in `RevMet 2.0` [Wernersson, 2003], after which we curated the sequences by eye, fixing obvious homopolymer errors and removing sequences with uncorrectable stop codons and those that failed to align well to the others, the latter two likely being 'Numts' (pseudogenes from nuclear insertions of mitochondrial sequences). This left us with 1520 OTUs.

In the final step, we read in the taxonomies of these OTUs and visually checked pairs of OTUs that had received very similar taxonomies (IDd to the same BOLDID) for which one OTU contained many reads and the other contained few. These are likely oversplit OTUs, and we removed the smaller of the OTUs. In rare cases, there are multiple OTUs that match to the same BOLDID, but one or more of them are only BLAST weak matches to that BOLDID and contain many reads, suggesting that these OTUs are true species for which reference sequences do not exist. Our bias throughout is to remove OTUs that could be artefactual splits of true OTUs, because these small OTUs will interfere with read mapping and do not add true diversity to the dataset. We were left with 1225 OTUs as the reference barcode set, and to this fasta file, we added the two spike-in COI sequences.

Read mapping with minimap2, samtools, and bedtools

We then used the newly constructed reference barcode dataset to detect species in each sample's shotgun reads. This is done by applying a commonly used tool from genomics known as a sequence alignment program, which maps individual Illumina reads against one or more reference sequences (usually a genome, but here the reference barcodes). Reference barcodes to which multiple Illumina reads are aligned are taken to be present in that sample, as long as the read mappings are (1) high quality (close match, low estimated error rate, map in the correct orientation) and (2) cover more than 50% of the barcode length, under the logic that if a species is truly in a sample, reads from the whole COI gene will be in the sample and will thus 'map' along the length of that species' barcode. These acceptance criteria were determined with experimental mock samples of known composition [Ji et al., 2020]. The output of mapping all samples individually to the reference barcodes is a sample x species table. After removing a few samples that were missing sample-identifying metadata or had no mapped reads to the spike-ins, we were left with 237 samples of the original 242, of which 121 were from sampling session 1 (July 2018).

We used `minimap2 2.17-r941` (Li 2018) in short-read mode (`minimap2 -ax sr`) to map the read pairs from each sample to the 1225 reference barcodes and the 2 spike-in sequences. We used `samtools 1.5` [Li, 2018] to sort, convert to bam format, exclude reads that were unmapped or mapped as secondary alignments and supplementary alignments, and include only 'proper-pair' read mappings (mapped in the correct orientation and at approximately the correct distance apart) at $\geq$ 48 'mapping quality' (MAPQ) (`samtools view sort -b -F 2308 -f 0x2 -q 48`).

$$MAPQ = -10log_{10}(\text{prob that mapping position is wrong})$$

We accepted $MAPQ \geq 48$ after inspection of the highly bimodal distribution of quality values, with most reads giving $MAPQ = 60$ (probability of error = 0.000001) or 0 (i.e. maps well to multiple locations). $MAPQ = 48$ corresponds to an error probability $\sim$ 0.000016. Informally, we have found that limiting quality to only the highest value, 60, has little effect on the results, whereas including low-quality mappings (`-q 1`) leads to more false-positive hits (data not shown). Read mapping data were output to `samtools idxstats` files.

The output for each sample is the number of mapped reads per OTU and spike-in that have passed the above filters. However, it is still possible for a barcode to receive false-positive mappings. Thus, we applied a second round of filtering. We expect that if a species is truly in a sample, reads from that sample will map *along the length* of that species' barcode, resulting in a high percentage coverage. In contrast, if reads map to just one location on a barcode, even at high MAPQ, the percentage coverage will be low, and we consider those mappings to be false-positive detections caused by that mapped portion of the barcode being very similar to a species that is in the sample but not in the reference database. We used `bedtools 2.29.2` [Quinlan and Hall, 2010] to calculate the number of overlapping reads at each position along the reference sequence (`genomecov -d`). The percent coverage is the fraction of positions in a barcode covered by one or more mapped reads. We kept only those species detections with percent coverage $\geq 50\%$, following recommendations from an experiment in Ji et al. [2020].

Sample X Species table creation

We imported the sample metadata and the samtools and bedtools outputs into `R 4.0.4` [R Core Team, 2022] for downstream processing into a sample x OTU table. After removing a few sites that had missing sample-identifying metadata or had no mapped reads to the spike-ins, we were left with 237 samples out of the original 242 (Table S S1M1M2). These samples represented two sampling sessions, of which 121 were in sampling Session 1 (July 2018) and 116 in Session 2 (August 2018). The 121 samples from Session 1 were distributed over 89 sites, of which 57 sites had 1 Malaise trap-sample and 32 sites had 2 samples. For this study, we used only the Session 1 samples. The two sessions only partially overlapped in species composition, meaning that it was not possible to test a Session 1 model on Session 2.

Environmental covariates

We used environmental covariates related to forest structure, vegetation reflectance and phenology, topography, anthropogenic features, and location to model arthropod incidence. We extracted the forest structure variables from lidar data collected from 2008 to 2016, consisting of 95$^{\text{th}}$ percentile canopy height, canopy cover above 2 and 4 m (calculated as the proportion of returns for a 30 m pixel above that height) and proportional area with canopy cover (calculated as the proportion of area with vegetation greater than 4 m) (Table 1S). These types of measures of canopy height and cover are correlated with field observations of forest structure in Pacific Northwest coniferous forests, such as mean diameter, canopy cover, and tree density [Kane et al., 2010]. We calculated vegetation indices from Landsat 8 images over the year, 2018, including Normalized Difference Vegetation Index (NDVI), Normalized Difference Moisture Index (NDMI), and Normalized Burn Ratio (NBR). From these, we calculated annual metrics of standard deviation, median, 5% and 95% percentiles over the year 2018, as well as using raw bands from a single cloudless image from 26/07/2018 (within 7 days of data collection). Both the proportion of canopy cover and annual Landsat metrics were calculated within the radii of 100, 250 and 500 m, given that vegetation structure at different spatial scales is known to drive arthropod biodiversity [Müller et al., 2014]. We created topographic predictors based on 1 m resolution bare-earth models from lidar ground returns, including elevation, slope, Eastness and Northness split from aspect, Topographic Position Index (TPI), Topographic Roughness Index (TRI) [Wilson et al., 2007], Topographic Wetness Index (TWI) [Metcalfe et al., 2018], and distance to streams, based on a vector stream network (`http://oregonexplorer.info`, accessed 24 Oct 2019). We used spatial data on anthropogenic activities to create predictors based on distance to nearest road, proportion of area logged within the last 100 and

40 years within radii of 250, 500 and 1000 m, and a categorical variable of inside or outside the boundary of the H.J. Andrews Experimental Forest. We used the `raster` and `sf` packages for `R` for all spatial analysis [Hijmans, 2022, Pebesma, 2018]. We mapped all 58 candidate environmental covariates (Table S-1) at 30 m resolution — either matching native resolution (e.g. Landsat), or aggregated from finer resolution data (e.g. lidar data), and projected them to the UTM 10N grid.

Statistical Analyses

Species inputs

For modelling, we converted the sequence-read-number OTU table to presence-absence (1/0), and we only included OTUs present at ≥ 6 sampling sites across the 121 samples. Our species dataset thus consisted of 190 OTUs in two classes, Insecta and Arachnida (Figure **??**).

Environmental covariates

To avoid collinearity, which would pose problems for the application of explainable AI [xAI, see below; Hooker et al., 2021], we iteratively calculated the Variance Inflation Factor [VIF; Zuur et al., 2007] on the 58 scaled candidate covariates, eliminating the highest scoring variable each time until all VIF values were < 8. The exception is that we forced the covariates elevation and inside/outside H.J. Andrews Forest to remain within the set of predictors irrespective of their VIF value, for a total of 29 predictors.

Joint Species Distribution Model

The general idea behind species distribution modelling is to ”predict a species' distribution”, using the species' observed incidences (presences and absences) and the combination of environmental-covariate values (i.e. the 29 covariates) in those points, to estimate the probability of species' incidences (i.e. to 'fit the model'). After model fitting, species in the rest of the sampling area, where environmental conditions are known but species' incidences are not, can be predicted, and the fitted model uses the environmental-covariate values to calculate the species' probability of presence. In this way, each species' distribution is predicted across continuous space, with varying degrees of accuracy.

We used the `R` package `sjSDM 0.1.6` [Pichler and Hartig, 2021], which is a JSDM that implements an integral approximation of multivariate probit models. sjSDM also includes a DNN (deep neural network) option to fit environmental covariates, which suits our dataset of many species with few data points and many covariates. We modeled the presence-absence data with a bino mial distribution (probit link) in the sjSDM framework. The species occurrence probabilities are described as a function of a three-layer DNN on the environmental covariates in addition to spatial coordinates to account for spatial auto-correlation and a species covariance matrix:

$$Z_{ij} = \beta_{0j} + DNN(X_{in}) + X_{s_i}\beta_{s_j} + MVN(0, \Sigma_{ij})$$

$$Y_{ij} = 1(Z_{ij} > 0),$$

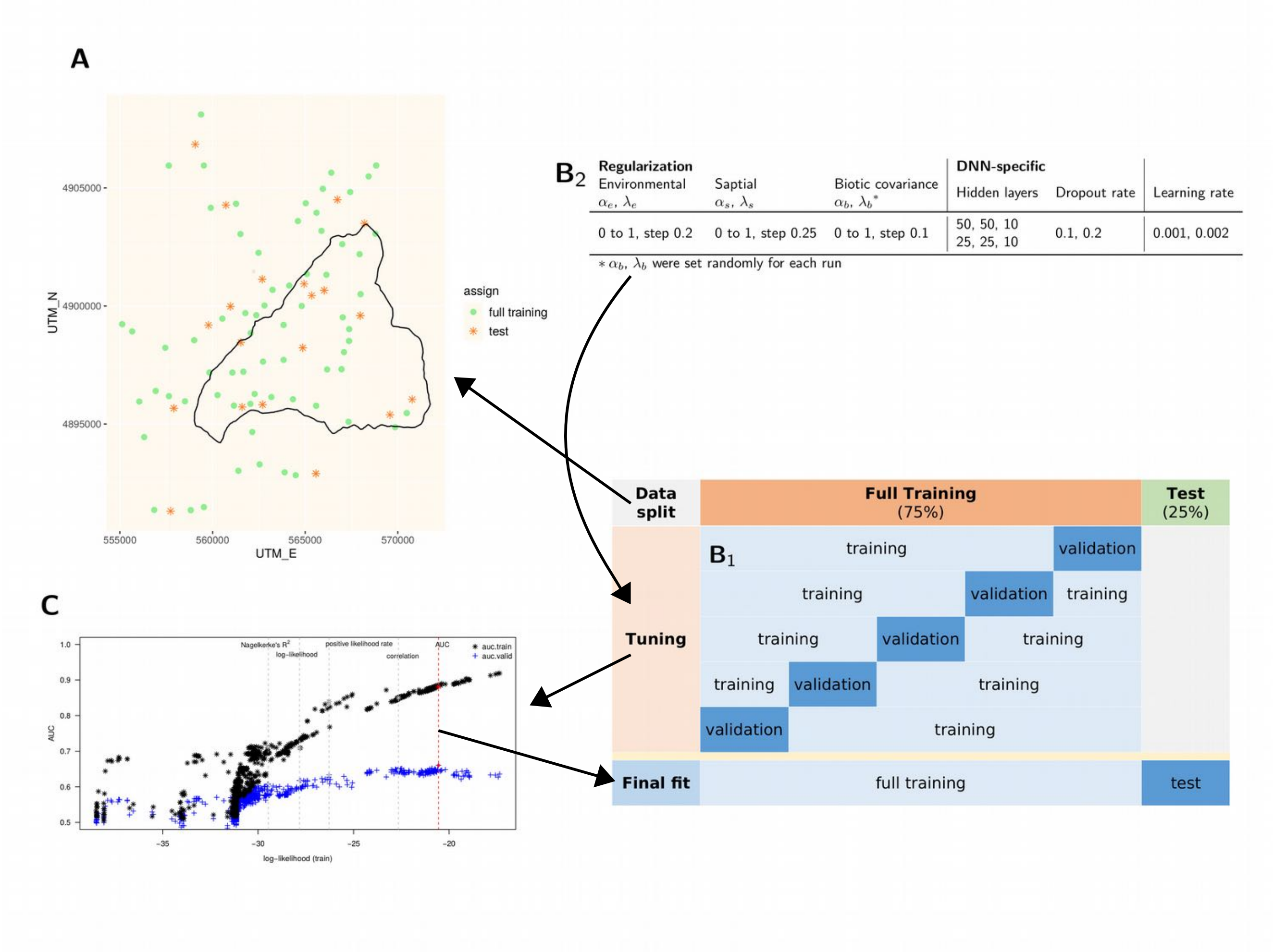
in which Z_{ij} is the occurrence probability of species j at sampling site i; Y_{ij} is the observed presence of species j at site i; X_{in} is the value of environmental covariate n in sampling site i; $X_{s_i}\beta_{s_j}$ is the spatial term, which includes the individual and interaction terms of two Universal Transverse Mercator variables (X_{s_i} is the coordinate variable for sampling site i, and β_{s_j} the coefficient of the coordinate variable for species j); MVN is the multivariate normal error representing the species correlation matrix.

Tuning and testing

The statistical challenge is to avoid overfitting, which is when the fitted model does a good job of predicting the species' incidences in the sampling points that were used to fit the model in the first place but does a bad job of predicting the species over the rest of the landscape. Overfitting is most likely to occur with species that have few presences (in our case, because we have zero-inflated data), with large numbers of environmental covariates, and when the model uses flexible mathematical functions to describe the relationships between environmental-covariates and species incidences. Unfortunately, all three of these conditions apply when trying to model arthropod fine-scale distributions. Many species are rare, there are many candidate remote-sensing covariates, and we expect that any relationships between remote-sensing-derived covariates and arthropod incidences will be indirect and thus complex, necessitating the use of flexible mathematical functions.

We randomly split the 121 data points from July 2018 into 75% training data ($n = 91$) and 25% test data ($n = 30$) (i.e. hold-out data), and we ensured that when two Malaise traps had been placed at the same site, they were assigned to the same split (Figure 1M).

We tuned nine hyperparameters of the sjSDM model with 5-fold cross-validation on the training data, also ensuring that data from pairs of traps placed at the same site were assigned to the same fold. During each round of tuning (i.e. each hyperparameter combination), five models were run. Four folds of the training data were used for training, the fifth was used for evaluation of the trained model (validation data), and the folds were rotated to produce five evaluations (Figure 1M). The nine hyperparameters consisted of the weighting between lasso and ridge regularization parameters ($\alpha_{e,s,b}$) and their strength ($\lambda_{e,s,b}$) for each of the environmental, spatial, and species covariance components, the dropout rate, the hidden structure for the DNN, and the learning rate of the model (Figure 1MC). We randomly selected 1000 combinations from the full tuning grid ($n = 7200$), and the lambda and alpha parameters of the environmental covariance were chosen randomly from their possible ranges (0 to 1; see Figure 1MC) for each round of tuning. We used six metrics to evaluate tuning performance: AUC (area under the receiver operating characteristic curve), positive likelihood ratio, Pearson's correlation coefficient, log-likelihood,



B_2

Regularization Environmental α_e, λ_e	Saptial α_s, λ_s	Biotic covariance α_b, λ_b*	**DNN-specific** Hidden layers	Dropout rate	Learning rate
0 to 1, step 0.2	0 to 1, step 0.25	0 to 1, step 0.1	50, 50, 10 25, 25, 10	0.1, 0.2	0.001, 0.002

$*$ α_b, λ_b were set randomly for each run

Figure 1M: Model tuning and training strategy. We obtained our final model by data splitting, tuning, and final fitting. *A*. We randomly split the 121 Malaise traps into test ($n = 30$) and training subsets (91). B_1. We then randomly split the training set into five parts for tuning via a 5-fold cross-validation. For all sets of splits, when a sampling site contained two Malaise traps, both traps were assigned to the same split. During each round of tuning (same hyperparameters combination), five models are run with one fold as the validation data and four folds as training. B_2. We randomly sampled 1000 rows from a tuning grid of all combinations of hyperparameters ($n = 7200$), and the performance of each tuning model was tested against the validation data. λ sets the overall strength of regularization, and α sets the relative weighting of ridge vs. lasso penalties. *C*. After finding the best combination of hyperparameters for the AUC (area under the ROC curve) performance metric, we fit the model to the full training data and tested the fitted model's predictive power against the test data. The black asterisks are the average AUC values for the training sets, and the blue crosses are the average for the validation sets.

True Skill Statistic (TSS), and Nagelkerke's R2 [Lawson et al., 2014, Wilkinson et al., 2021, see Supplementary Information]. For each hyperparameter combination, metric, and fold, we recorded the explanatory performance (on the training data) and the predictive performance (on the validation data), and we averaged the five folds as the evaluation for each hyperparameter combination. We trained a final model on the full training dataset using the best hyperparameter combination as judged by the highest predictive AUC performance. We recorded the final model's explanatory AUC from the training dataset, and finally we used the test dataset to evaluate the predictive AUC performances of the final model (Figure 1MB_1). The final models chosen by the other four performance metrics behaved similarly (see Figure 1S).

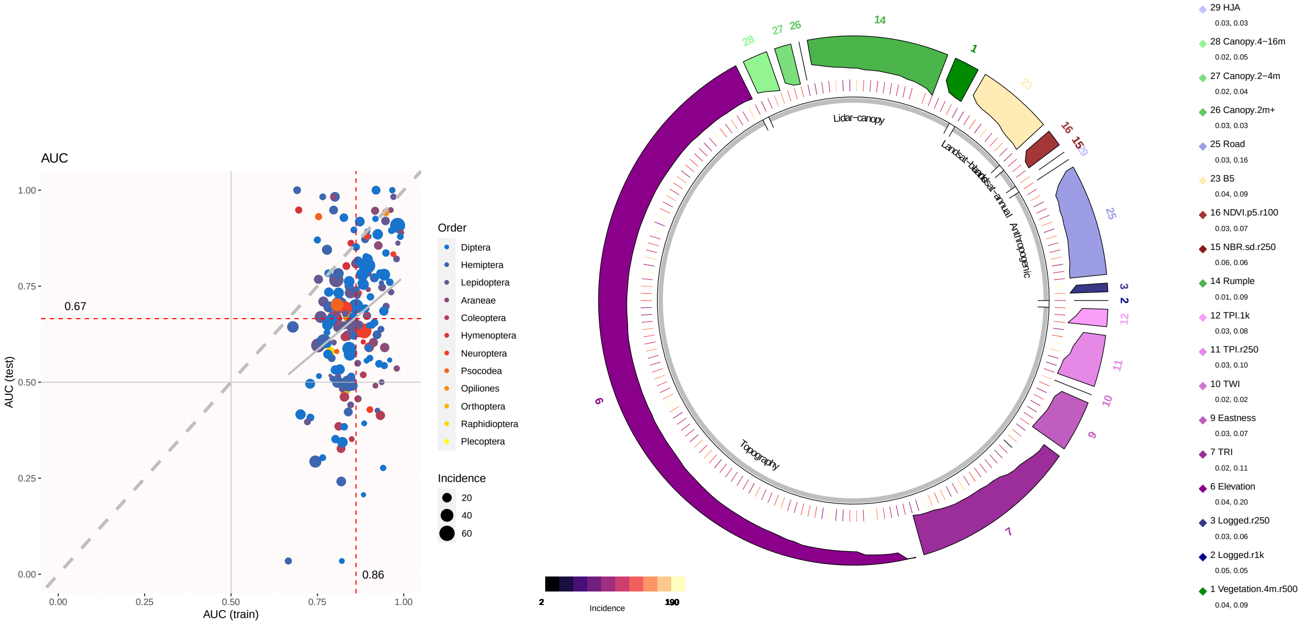
Variable importance with explainable AI (xAI)

To gain insight into the importances of the environmental covariates in our DNN, we analyzed variable importance using permutation and Friedman's H statistics, as implemented in the R package `flashlight 0.8.0` [Mayer, 2021]. The permutation statistic evaluates the overall importance of each variable, i.e. the decrease in performance (here, AUC, see below) when permuting the values of that variable [Fisher et al., 2019], and the Friedman's H statistic evaluates the overall interaction strength of the variables, i.e. the strength of the non-additive effect of one variable on the full model, based on a partial dependence function [Friedman and Popescu, 2008]. We omitted the spatial component when calculating variable importance.

We calculated these metrics of xAI based on the explanatory performance of the sjSDM model, and the AUC performance matrix was used. The variable importance was calculated by permuting all data points of the environmental covariates over six repetitions to ensure a stable result. Afterwards, we chose the ten most important covariates based on the resulting variable importance for each species to conduct the unnormalized H-statistics. The unnormalized H-statistics were chosen to ensure a fair comparison between variables. The H-statistic was calculated using all the data points as well.

Prediction and visualisation of species distributions

Using the final model, we show three examples of how to visualize species predictions. Firstly, we used the final model to predict the distributions of those species with predictive AUC > 0.7. To avoid extrapolation [Norberg et al., 2019], we restricted predictions to a 1 km buffered, convex hull around all sample sites, edited manually to avoid suburban areas in the southern extreme of the study area. Further, all predictors within this area were restricted, or 'clamped', to lie within the range of predictor values across all sample points, that is, predictors above or below this range were given the maximum or minimum value from across the sample points, respectively [Anderson and Raza, 2010]. Given the stochasticity inherent in sjSDM predictions [Pichler and Hartig, 2021], each species' prediction used the average of five separate prediction runs. We created binary species distributions maps



(a)

(b)

Figure 2M: Model performance and environmental-covariate importance. (a). Explanatory AUC (range 0.67-1, mean 0.86, median 0.86) and predictive AUC (range 0.03-1, mean 0.67, median 0.67) of the final model. Each point is one OTU. Color indicates taxonomic class (order), and point size indicates incidence (number of Malaise traps in which the OTU was detected). Predictive AUC value is not explained by incidence (linear model, p = 0.93, R^2 = $4.5e-05$). The dashed gray line is the 1:1 line, and the solid gray line is a fitted linear regression. (b). Most important explanatory environmental covariate for each OTU, as determined by xAI (see Variable importance with explainable AI). Tick marks indicate each OTU's incidence, color bands indicate individual covariates, and gray bands indicate logical covariate groupings (Table 1S). Elevation (variable 6) and Topographic Roughness Index (variable 7) are the most important individual environmental covariates for the most OTUs, and the six variables in the topography group are the most important as a group. The heights of the colour bars are scaled to the permutation importance for that OTU.

by applying a 0.5 threshold on the occurrence probability values, and summed these to create a species richness map. We acknowledge that a common threshold for all species is not ideal, but no further analysis is performed with the binary maps.

Secondly, to map community similarity across the study area, we ordinated species predictions on two dimensions using T-SNE (t-Distributed Stochastic Neighbor Embedding) and mapped the two resulting ordination axes. T-SNE is a dimension-reduction technique where high-dimensional distances between data points are converted into conditional probabilities that represent similarities [van der Maaten and Hinton, 2008]. The R implementation [Krijthe, 2015] uses the Barnes-Hut approximation to increase performance with large data sets. The perplexity parameter, which controls the number of points available within the neighborhood, was set at 50.

Finally, after applying the final model to the test dataset, we identified 76 species that had moderate to high predictive performance. We used the fitted model and the environmental-covariates to predict the probability of each species' incidence in each grid cells in the study area ('filling in the blanks' between the sampling points). The output is 76 individual and continuous species distribution maps, which we stacked to carry out three landscape analyses. First, we counted the number of species predicted to be present (probability of presence $\geq 50\%$) in each grid square to produce a species richness map. Second, we carried out a dimension-reduction analysis, also known as ordination, using the T-SNE method [van der Maaten and Hinton, 2008, Krijthe, 2015] to summarise species compositional change across the landscape. Pixels that have similar species compositions receive similar T-SNE values, which can be visualised. Third, we calculated Baisero et al. [2022] site-irreplaceability index for every pixel. This index is the probability that loss of that pixel would prevent achieving the conservation target for at least one of the 76 species, where the conservation target is set to be 50% of the species' total incidence.

Thirdly, we calculated the Baisero et al. [2022] site-irreplaceability index (β) per pixel across the study area as the combined probability that a site is irreplaceable for at least one OTU. The beta index combines species-level irreplaceability indices, alpha, at each site, measured as proximity-based metrics of how close a site is to being required to achieve a conservation target for a particular species. We used a value of 50% of each species' total incidence across the study area as our conservation target.

Finally, we carried out post-hoc analyses by plotting site irreplaceability, composition (T-SNE), and species richness against elevation, old-growth structural index [Davis et al., 2015], and inside/outside HJA. We consider these analyses to be post-hoc because we are applying them to the predicted species distributions, which we viewed before analysis. Thus, we consider these analyses to be hypothesis-generating exercises for future studies.

Caveats

Irreplaceability

We used Baisero et al.'s (2022) method to calculate site irreplaceability. Two advantages are that it is fast to calculate and is stable to changes in the grid system and in the addition or subtraction of species from the dataset, unlike the alternative method of using selection frequency from the outcome of a systematic conservation planning (SCP) algorithm, which additionally must assume that the sites selected by an SCP run are optimal. Baisero et al.'s (2022) site-irreplaceability value is one minus the probability that a site is replaceable for all species in that site. A value of 0 means that a site's loss would still allow the conservation target of every species in that site to be met using other sites in the landscape, where a target is the proportion of a species' range that is designated for protection. Thus, sites with higher irreplaceability values are characterised by higher numbers of species with high targets and/or small ranges. The latter reason is why lower elevations, the riverine basin (including the southern edge, which borders a river), and plantations are given high irreplaceability values (Figure **??** B), since these habitat types (and their associated species) cover a smaller proportion of the total landscape, and thus any species limited to them needs those sites protected for their conservation targets to be met (Figure **??** A). It is important to keep in mind that any measure of site irreplaceability can only compare the sites *within* an analysed landscape, meaning that a small pine plantation in a tropical rainforest would be scored high on irreplaceability if it contained pine specialist arthropods. For such situations, known widespread and common species can be given low conservation targets, and artefactually rare habitats (the plantation in a rainforest) can be masked from analysis. For instance, we repeated the site-irreplaceability analysis after masking plantations, since recently logged forest characterises most of the Oregon forest landscape outside the H.J. Andrews Experimental Forest. Without plantations, areas near streams increased in irreplaceability value (Figure 8S).

False-negative error

Despite detecting 1225 OTUs across the whole dataset, ultimately, only 76 OTUs had enough detections to be modelled and mapped. An independent analysis of this dataset has estimated that even the 50 most prevalent species have only a $\sim 50\%$ probability of being detected when they are truly at the sampling points [Diana et al., 2022]. Consequently, we infer that many species absences are false negatives, which biases species prevalences and environmental-covariate effect sizes downwards. To increase the number of species that can be modelled, we make four recommendations:

1. Per sample, increase DNA-sequencing depth and/or increase the concentration of DNA barcode sequences using hybridisation or physical PCR [e.g. Liu et al., 2016, Yang et al., 2021].

2. Increase the number of sampling points.

3. Take multiple replicates per sampling point. Roughly, the per-bulk-arthropod-sample cost of the mitogenome mapping protocol is $\sim$ US\$250, and commercial bulk-sample metabarcoding prices (i.e. physical PCR) range from US\$100 to \$350 per sample. Two traps per 89 sites would cost $\$17,800$ to $\$62,300$ total, or \$79 to \$277 per km^2. Using multiple traps per site directly reduces the rate of false negatives and provides the option of combining occupancy correction and JSDMs [Doser et al., 2022, Tobler et al., 2019, Diana et al., 2022] to account for false-negative error.

4. Change the trapping method. Malaise traps seem especially prone to false-negative error [Steinke et al., 2021]. An alternative is pitfall traps, for which it is cheap to increase trapping effectiveness [by adding cups and guidance barriers, Boetzl et al., 2018].

Errors in environmental covariates

We used both LANDSAT and multiple lidar datasets collected from 2008-2016 to generate predictors for species data collected in 2018, following successful use of Earth Observation data for biodiversity mapping in other studies [Bae et al., 2019, Galbraith et al., 2015, Lin et al., 2021, Müller et al., 2009, Müller and Brandl, 2009]. The temporal mismatch between lidar and field data might introduce some errors [Gatziolis and Andersen, 2008] if major vegetation changes had occurred between acquisitions (e.g. tree mortality), but in most cases, we expect forests to change slowly [Zald et al., 2014]. Differences in lidar collection specifications, especially lidar pulse density, which varied by roughly a factor of two, might also introduce artifacts if some metrics are particularly sensitive [e.g. Görgens et al., 2015] or are simply hard to reproduce [e.g. metrics based on lidar intensity, Bater et al., 2011]. That said, canopy height and cover metrics used in this study are likely relatively stable across acquisitions, and the LANDSAT data used in our model were collected during the sampling period, with a view to capturing species' niche axes such as vegetation phenology, habitat type and condition [Leitão and Santos, 2019].

Choice of JSDM software and interpretation

Our choice of sjSDM over other JSDM software packages was largely dictated by sjSDM's much faster runtimes while exhibiting predictive performance levels that match other packages [Pichler and Hartig, 2021]. sjSDM also uniquely provides the option to use a combination of regularization and a deep neural network for model fitting, which is appropriate for situations with large numbers of environmental covariates, such as our use of remote-sensing layers, and where the focus is on the predictive power of a model. To compare the effect of using a DNN, we reran the sjSDM model with the same setup but linear in the environmental part. Both explanatory and predictive power of the linear version are not as high as the DNN model (Figure 9S). A DNN fitting procedure thus appears to be useful for disentangling complex relationships between remote-sensing-derived environmental covariates and community data.

Joint species distribution models are distinguished by estimating not only species responses to environmental covariates (as in all species distribution models) but also by estimating correlations between all species pairs while accounting for environmental responses. These residual species associations can be interpreted as the effect of unmeasured environmental covariates and/or the effect of biotic interactions, such as competition or facilitation [Ovaskainen et al., 2017, Pollock et al., 2014, Warton et al., 2015]. It has proven difficult to distinguish between the two in practice [Dormann et al., 2018, König et al., 2021, Poggiato et al., 2021, Zurell et al., 2018, Hartig et al., 2023], and in this study, we are agnostic as to the interpretation of residual species correlations.

References

Robert P. Anderson and Ali Raza. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela: Effect of study region on models of distributions. *Journal of Biogeography*, 37(7):1378–1393, April 2010. ISSN 03050270, 13652699. doi: 10.1111/j.1365-2699.2010.02290.x. URL `https://onlinelibrary.wiley.`

com/doi/10.1111/j.1365-2699.2010.02290.x.

Soyeon Bae, Shaun R. Levick, Lea Heidrich, Paul Magdon, Benjamin F. Leutner, Stephan Wöllauer, Alla Serebryanyk, Thomas Nauss, Peter Krzystek, Martin M. Gossner, Peter Schall, Christoph Heibl, Claus Bässler, Inken Doerfler, Ernst-Detlef Schulze, Franz-Sebastian Krah, Heike Culmsee, Kirsten Jung, Marco Heurich, Markus Fischer, Sebastian Seibold, Simon Thorn, Tobias Gerlach, Torsten Hothorn, Wolfgang W. Weisser, and Jörg Müller. Radar vision in the mapping of forest biodiversity from space. *Nature Communications*, 10(1):4757, December 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-12737-x. URL http://www.nature.com/articles/s41467-019-12737-x.

Daniele Baisero, Richard Schuster, and Andrew J. Plumptre. Redefining and mapping global irreplaceability. *Conservation Biology*, 36(2), April 2022. ISSN 0888-8892, 1523-1739. doi: 10.1111/cobi.13806. URL https://onlinelibrary.wiley.com/doi/10.1111/cobi.13806.

Christopher W. Bater, Michael A. Wulder, Nicholas C. Coops, Ross F. Nelson, Thomas Hilker, and Erik Nasset. Stability of Sample-Based Scanning-LiDAR-Derived Vegetation Metrics for Forest Monitoring. *IEEE Transactions on Geoscience and Remote Sensing*, 49(6):2385–2392, June 2011. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2010.2099232. URL http://ieeexplore.ieee.org/document/5696751/.

Jeremy Biggs, Naomi Ewald, Alice Valentini, Coline Gaboriaud, Tony Dejean, Richard A. Griffiths, Jim Foster, John W. Wilkinson, Andy Arnell, Peter Brotherton, Penny Williams, and Francesca Dunn. Using edna to develop a national citizen science-based monitoring programme for the great crested newt (triturus cristatus). *Biological Conservation*, 183:19–28, Mar 2015. ISSN 00063207. doi: 10.1016/j.biocon.2014.11.029.

Fabian A. Boetzl, Elena Ries, Gudrun Schneider, and Jochen Krauss. It's a matter of design—how pitfall trap design affects trap samples and possible predictions. *PeerJ*, 6:e5078, June 2018. ISSN 2167-8359. doi: 10.7717/peerj.5078. URL https://peerj.com/articles/5078.

Raymond J. Davis, Janet L. Ohmann, Robert E. Kennedy, Warren B. Cohen, Matthew J. Gregory, Zhiqiang Yang, Heather M. Roberts, Andrew N. Gray, and Thomas A. Spies. Northwest Forest Plan–the first 20 years (1994-2013): status and trends of late-successional and old-growth forests. Technical Report PNW-GTR-911, U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, OR, 2015. URL https://www.fs.usda.gov/treesearch/pubs/50060.

Alex Diana, Eleni Matechou, Jim Griffin, Douglas W. Yu, Mingjie Luo, Marie Tosa, Alex Bush, and Richard Griffiths. eDNAPlus: A unifying modelling framework for dna-based biodiversity monitoring. (arXiv:2211.12213), Nov 2022. URL http://arxiv.org/abs/2211.12213. arXiv:2211.12213 [stat].

Carsten F. Dormann, Maria Bobrowski, D. Matthias Dehling, David J. Harris, Florian Hartig, Heike Lischke, Marco D. Moretti, Jörn Pagel, Stefan Pinkert, Matthias Schleuning, Susanne I. Schmidt, Christine S. Sheppard, Manuel J. Steinbauer, Dirk Zeuss, and Casper Kraan. Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. *Global Ecology and Biogeography*, 27(9):1004–1016, September 2018. ISSN 1466822X. doi: 10.1111/geb.12759. URL https://onlinelibrary.wiley.com/doi/10.1111/geb.12759.

Jeffrey W. Doser, Andrew O. Finley, Marc Kéry, and Elise F. Zipkin. spoccupancy: An r package for single-species, multi-species, and integrated spatial occupancy models. *Methods in Ecology and Evolution*, 13(8):1670–1678, 2022. ISSN 2041-210X. doi: 10.1111/2041-210X.13897.

Vasco Elbrecht, Thomas W.A. Braukmann, Natalia V. Ivanova, Sean W.J. Prosser, Mehrdad Hajibabaei, Michael Wright, Evgeny V. Zakharov, Paul D.N. Hebert, and Dirk Steinke. Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, 7:e7745, October 2019. ISSN 2167-8359. doi: 10.7717/peerj.7745. URL https://peerj.com/articles/7745.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously, December 2019. URL http://arxiv.org/abs/1801.01489. Number: arXiv:1801.01489 arXiv:1801.01489 [stat].

Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), September 2008. ISSN 1932-6157. doi: 10.1214/07-AOAS148. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/Predictive-learning-via-rule-ensembles/10.1214/07-AOAS148.full.

Sara M. Galbraith, L. A. Vierling, and N. A. Bosque-Pérez. Remote Sensing and Ecosystem Services: Current Status and Future Opportunities for the Study of Bees and Pollination-Related Services. *Current Forestry Reports*, 1(4):261–274, December 2015. ISSN 2198-6436. doi: 10.1007/s40725-015-0024-6. URL http://link.springer.com/10.1007/s40725-015-0024-6.

Demetrios Gatziolis and Hans-Erik. Andersen. A guide to LIDAR data acquisition and processing for the forests of the Pacific Northwest. Technical Report PNW-GTR-768, U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, OR, 2008. URL https://www.fs.usda.gov/treesearch/pubs/30652.

Paul Greenfield, Nai Tran-Dinh, and David Midgley. Kelpie: generating full-length 'amplicons' from whole-metagenome datasets. *PeerJ*, 6:e6174, January 2019. ISSN 2167-8359. doi: 10.7717/peerj.6174. URL https://peerj.com/articles/6174.

Eric Bastos Görgens, Petteri Packalen, André Gracioso Peres da Silva, Clayton Alcarde Alvares, Otavio Camargo Campoe, José Luiz Stape, and Luiz Carlos Estraviz Rodriguez. Stand volume models based on stable metrics as from multiple ALS acquisitions in Eucalyptus plantations. *Annals of Forest Science*, 72(4):489–498, June

2015. ISSN 1286-4560, 1297-966X. doi: 10.1007/s13595-015-0457-x. URL `http://link.springer.com/10.1007/s13595-015-0457-x`.

Florian Hartig, Nerea Abrego, Alex Bush, Jonathan M. Chase, Gurutzeta Guillera-Arroita, Mathew A. Leibold, Otso Ovaskainen, Loïc Pellissier, Maximilian Pichler, Giovanni Poggiato, Laura Pollock, Sara Si-Moussi, Wilfried Thuiller, Duarte S. Viana, David Warton, Damaris Zurell, and Douglas W. Yu. Novel community data – properties and prospects. 2023.

Robert J. Hijmans. *raster: Geographic Data Analysis and Modeling*. 2022. URL `https://CRAN.R-project.org/package=raster`.

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):82, November 2021. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-021-10057-z. URL `https://link.springer.com/10.1007/s11222-021-10057-z`.

Yinqiu Ji, Tea Huotari, Tomas Roslin, Niels Martin Schmidt, Jiaxin Wang, Douglas W. Yu, and Otso Ovaskainen. SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Molecular Ecology Resources*, 20(1):256–267, January 2020. ISSN 1755-098X, 1755-0998. doi: 10.1111/1755-0998.13057. URL `https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13057`.

Van R. Kane, Robert J. McGaughey, Jonathan D. Bakker, Rolf F. Gersonde, James A. Lutz, and Jerry F. Franklin. Comparisons between field- and LiDAR-based measures of stand structural complexity. *Canadian Journal of Forest Research*, 40(4):761–773, April 2010. ISSN 0045-5067, 1208-6037. doi: 10.1139/X10-024. URL `http://www.nrcresearchpress.com/doi/10.1139/X10-024`.

Jesse H. Krijthe. *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*, 2015. URL `https://github.com/jkrijthe/Rtsne`. R package version 0.15.

Christian König, Rafael O. Wüest, Catherine H. Graham, Dirk Nikolaus Karger, Thomas Sattler, Niklaus E. Zimmermann, and Damaris Zurell. Scale dependency of joint species distribution models challenges interpretation of biotic interactions. *Journal of Biogeography*, 48(7):1541–1551, July 2021. ISSN 0305-0270, 1365-2699. doi: 10.1111/jbi.14106. URL `https://onlinelibrary.wiley.com/doi/10.1111/jbi.14106`.

Callum R. Lawson, Jenny A. Hodgson, Robert J. Wilson, and Shane A. Richards. Prevalence, thresholds and the performance of presence-absence models. *Methods in Ecology and Evolution*, 5(1):54–64, January 2014. ISSN 2041210X. doi: 10.1111/2041-210X.12123. URL `https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12123`.

Pedro J. Leitão and Maria J. Santos. Improving Models of Species Ecological Niches: A Remote Sensing Overview. *Frontiers in Ecology and Evolution*, 7:9, January 2019. ISSN 2296-701X. doi: 10.3389/fevo.2019.00009. URL `https://www.frontiersin.org/article/10.3389/fevo.2019.00009/full`.

Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty191. URL `https://academic.oup.com/bioinformatics/article/34/18/3094/4994778`.

Meixi Lin, Ariel Levi Simons, Ryan J. Harrigan, Emily E. Curd, Fabian D. Schneider, Dannise V. Ruiz-Ramos, Zack Gold, Melisa G. Osborne, Sabrina Shirazi, Teia M. Schweizer, Tiara N. Moore, Emma A. Fox, Rachel Turba, Ana E. Garcia-Vedrenne, Sarah K. Helman, Kelsi Rutledge, Maura Palacios Mejia, Onny Marwayana, Miroslava N. Munguia Ramos, Regina Wetzer, N. Dean Pentcheff, Emily Jane McTavish, Michael N. Dawson, Beth Shapiro, Robert K. Wayne, and Rachel S. Meyer. Landscape analyses using eDNA metabarcoding and Earth observation predict community biodiversity in California. *Ecological Applications*, 31(6):e02379, September 2021. ISSN 1051-0761, 1939-5582. doi: 10.1002/eap.2379. URL `https://onlinelibrary.wiley.com/doi/10.1002/eap.2379`.

Shanlin Liu, Xin Wang, Lin Xie, Meihua Tan, Zhenyu Li, Xu Su, Hao Zhang, Bernhard Misof, Karl M. Kjer, Min Tang, Oliver Niehuis, Hui Jiang, and Xin Zhou. Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, 16(2):470–479, March 2016. ISSN 1755-098X, 1755-0998. doi: 10.1111/1755-0998.12472. URL `https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.12472`.

Michael Mayer. *flashlight: Shed Light on Black Box Machine Learning Models*, 2021. URL `https://github.com/mayer79/flashlight`. R package version 0.8.0.

Peter Metcalfe, Keith Beven, and Jim Freer. *dynatopmodel: Implementation of the Dynamic TOPMODEL Hydrological Model*. 2018. URL `https://CRAN.R-project.org/package=dynatopmodel`.

Jörg Müller and Roland Brandl. Assessing biodiversity by remote sensing in mountainous terrain: the potential of LiDAR to predict forest beetle assemblages. *Journal of Applied Ecology*, 46(4):897–905, August 2009. ISSN 00218901, 13652664. doi: 10.1111/j.1365-2664.2009.01677.x. URL `https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2664.2009.01677.x`.

Jörg Müller, Christoph Moning, Claus Bässler, Marco Heurich, and Roland Brandl. Using airborne laser scanning to model potential abundance and assemblages of forest passerines. *Basic and Applied Ecology*, 10(7):671–681, October 2009. ISSN 14391791. doi: 10.1016/j.baae.2009.03.004. URL `https://linkinghub.elsevier.com/retrieve/pii/S1439179109000280`.

Jörg Müller, Soyeon Bae, Juliane Röder, Anne Chao, and Raphael K. Didham. Airborne LiDAR reveals context dependence in the effects of canopy architecture on arthropod diversity. *Forest Ecology and Management*, 312:129–137, January 2014. ISSN 03781127. doi: 10.1016/j.foreco.2013.10.014. URL https://linkinghub.elsevier.com/retrieve/pii/S0378112713006816.

Natural England. *A Framework For District Licensing Of Development Affecting Great Crested Newts*. Number TIN176. Jul 2019. URL https://publications.naturalengland.org.uk/publication/5106496688095232. ISBN 978-1-78354-536-0.

Anna Norberg, Nerea Abrego, F. Guillaume Blanchet, Frederick R. Adler, Barbara J. Anderson, Jani Anttila, Miguel B. Araújo, Tad Dallas, David Dunson, Jane Elith, Scott D. Foster, Richard Fox, Janet Franklin, William Godsoe, Antoine Guisan, Bob O'Hara, Nicole A. Hill, Robert D. Holt, Francis K. C. Hui, Magne Husby, John Atle Kålås, Aleksi Lehikoinen, Miska Luoto, Heidi K. Mod, Graeme Newell, Ian Renner, Tomas Roslin, Janne Soininen, Wilfried Thuiller, Jarno Vanhatalo, David Warton, Matt White, Niklaus E. Zimmermann, Dominique Gravel, and Otso Ovaskainen. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3), August 2019. ISSN 0012-9615, 1557-7015. doi: 10.1002/ecm.1370. URL https://onlinelibrary.wiley.com/doi/10.1002/ecm.1370.

Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, and Nerea Abrego. How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5):561–576, May 2017. ISSN 1461023X. doi: 10.1111/ele.12757. URL https://onlinelibrary.wiley.com/doi/10.1111/ele.12757.

Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439, 2018. ISSN 2073-4859. doi: 10.32614/RJ-2018-009. URL https://journal.r-project.org/archive/2018/RJ-2018-009/index.html.

Maximilian Pichler and Florian Hartig. A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods in Ecology and Evolution*, 12(11):2159–2173, November 2021. ISSN 2041-210X, 2041-210X. doi: 10.1111/2041-210X.13687. URL https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13687.

Giovanni Poggiato, Tamara Münkemüller, Daria Bystrova, Julyan Arbel, James S. Clark, and Wilfried Thuiller. On the interpretations of joint modeling in community ecology. *Trends in Ecology & Evolution*, 36(5):391–401, May 2021. ISSN 01695347. doi: 10.1016/j.tree.2021.01.002.

Laura J. Pollock, Reid Tingley, William K. Morris, Nick Golding, Robert B. O'Hara, Kirsten M. Parris, Peter A. Vesk, and Michael A. McCarthy. Understanding co-occurrence by modelling species simultaneously with a Joint

Species Distribution Model (<span style="font-variant:small-caps;">JSDM</span>). *Methods in Ecology and Evolution*, 5(5):397–406, May 2014. ISSN 2041-210X, 2041-210X. doi: 10.1111/2041-210X.12180. URL https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12180.

Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btq033. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033.

R Core Team. R: A Language and Environment for Statistical Computing, 2022. URL https://www.R-project.org/.

Dirk Steinke, Thomas WA Braukmann, Laura Manerus, Allan Woodhouse, and Vasco Elbrecht. Effects of Malaise trap spacing on species richness and composition of terrestrial arthropod bulk samples. *Metabarcoding and Metagenomics*, 5:e59201, April 2021. ISSN 2534-9708. doi: 10.3897/mbmg.5.59201. URL https://mbmg.pensoft.net/article/59201/.

Mathias W. Tobler, Marc Kéry, Francis K. C. Hui, Gurutzeta Guillera-Arroita, Peter Knaus, and Thomas Sattler. Joint species distribution models with species correlations and imperfect detection. *Ecology*, 100(8), August 2019. ISSN 0012-9658, 1939-9170. doi: 10.1002/ecy.2754. URL https://onlinelibrary.wiley.com/doi/10.1002/ecy.2754.

Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

David I. Warton, F. Guillaume Blanchet, Robert B. O'Hara, Otso Ovaskainen, Sara Taskinen, Steven C. Walker, and Francis K.C. Hui. So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, 30(12):766–779, December 2015. ISSN 01695347. doi: 10.1016/j.tree.2015.09.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0169534715002402.

R. Wernersson. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Research*, 31(13):3537–3539, July 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg609. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg609.

David P. Wilkinson, Nick Golding, Gurutzeta Guillera-Arroita, Reid Tingley, and Michael A. McCarthy. Defining and evaluating predictions of joint species distribution models. *Methods in Ecology and Evolution*, 12(3):394–404, March 2021. ISSN 2041-210X, 2041-210X. doi: 10.1111/2041-210X.13518. URL https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13518.

Margaret F. J. Wilson, Brian O'Connell, Colin Brown, Janine C. Guinan, and Anthony J. Grehan. Multiscale Terrain Analysis of Multibeam Bathymetry Data for Habitat Mapping on the Continental Slope. *Marine Geodesy*, 30(1-2):3–35, May 2007. ISSN 0149-0419, 1521-060X. doi: 10.1080/01490410701295962. URL http://www.tandfonline.com/doi/abs/10.1080/01490410701295962.

Helena K. Wirta, Paul D. N. Hebert, Riikka Kaartinen, Sean W. Prosser, Gergely Várkonyi, and Tomas Roslin. Complementary molecular information changes our perception of food web structure. *Proceedings of the National Academy of Sciences*, 111(5):1885–1890, February 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1316990111. URL https://pnas.org/doi/full/10.1073/pnas.1316990111.

Chunyan Yang, Kristine Bohmann, Xiaoyang Wang, Wang Cai, Nathan Wales, Zhaoli Ding, Shyam Gopalakrishnan, and Douglas W. Yu. Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *Methods in Ecology and Evolution*, 12(7):1252–1264, July 2021. ISSN 2041-210X, 2041-210X. doi: 10.1111/2041-210X.13602. URL https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13602.

Harold S.J. Zald, Janet L. Ohmann, Heather M. Roberts, Matthew J. Gregory, Emilie B. Henderson, Robert J. McGaughey, and Justin Braaten. Influence of lidar, Landsat imagery, disturbance history, plot location accuracy, and plot size on accuracy of imputation maps of forest composition and structure. *Remote Sensing of Environment*, 143:26–38, March 2014. ISSN 00344257. doi: 10.1016/j.rse.2013.12.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0034425713004537.

Damaris Zurell, Laura J. Pollock, and Wilfried Thuiller. Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography*, 41(11):1812–1819, November 2018. ISSN 09067590. doi: 10.1111/ecog.03315. URL https://onlinelibrary.wiley.com/doi/10.1111/ecog.03315.

Alain F. Zuur, Elena N. Ieno, and Graham M. Smith. *Analysing ecological data*. Statistics for biology and health. Springer, New York, NY, 2007. ISBN 978-0-387-45972-1 978-0-387-45967-7.