PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

CC BY

Research

Cite this article: Li Y *et al*. 2024 Combining environmental DNA and remote sensing for efficient, fine-scale mapping of arthropod biodiversity. *Phil. Trans. R. Soc. B* **379**: 20230123.
https://doi.org/10.1098/rstb.2023.0123

Received: 31 August 2023
Accepted: 31 January 2024

One contribution of 23 to a theme issue 'Towards a toolkit for global insect biodiversity monitoring'.

Subject Areas:
ecology, environmental science

Keywords:
environmental DNA, Earth observation, biodiversity indices, systematic conservation planning, forestry, machine learning

Author for correspondence:
Douglas W. Yu
e-mail: dougwyu@mac.com

Downloaded from https://royalsocietypublishing.org/ on 06 May 2024

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

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

[1]Yunnan Key Laboratory of Biodiversity and Ecological Security of Gaoligong Mountain, State Key Laboratory of Genetic Resources and Evolution, and [2]Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, People's Republic of China
[3]Faculty of Biology, University of Duisburg-Essen, Essen 45141, Germany
[4]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR47TJ, UK
[5]Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State University, Corvallis, OR 97331, USA
[6]Kunming College of Life Sciences, University of Chinese Academy of Sciences, Kunming, People's Republic of China
[7]Pacific Northwest Research Station, U.S. Department of Agriculture Forest Service, Corvallis, OR 97331, USA
[8]CSIRO Energy, Lindfield, New South Wales, Australia
[9]School of Biological Sciences, Macquarie University, Sydney, Australia
[10]Theoretical Ecology, University of Regensburg, Regensburg, Germany
[11]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming Yunnan 650223, People's Republic of China

MIT, 0000-0002-9020-5098; DWY, 0000-0001-8551-5609

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 information 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 using a deep neural net in 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 combine the maps to visualize 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. With this 'sideways biodiversity modelling' method, we demonstrate the feasibility of biodiversity mapping at sufficient spatial resolution to inform local management choices, while also being efficient enough to scale up to thousands of square kilometres.

This article is part of the theme issue 'Towards a toolkit for global insect biodiversity monitoring'.

1. Introduction

Arthropods contribute in numerous ways to ecosystem functioning [1] but are understudied relative to vertebrates and plants [2]. This taxonomic bias

†Current address: School of Geography, Geology and the Environment, Keele University, Staffordshire, ST5 5BG, UK.
‡Co-first authors.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.7151335.

© 2024 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

THE ROYAL SOCIETY PUBLISHING

Downloaded from https://royalsocietypublishing.org/ on 06 May 2024

undermines the validity of conservation decisions when the effects of change in climate, land use and land cover differ across taxa [3,4]. 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 [5,6]. Another logistical advantage is that arthropod community structure is correlated with vegetation structure [7,8], and since vegetation can be measured remotely at large spatial scale via airborne and spaceborne sensors [9], remote imagery could also provide large-spatial-scale information on arthropods. In fact, it is already known that spaceborne synthetic aperture radar, and airborne light detection and ranging (LiDAR) imagery of fine-scale forest structure can predict the distributions of entomofauna and avifauna [10–13].

(a) Successful governance of the biodiversity commons

Arthropod conservation should be seen in the wider context of efficient biodiversity governance. Dietz *et al.*'s [14] framework for the successful governance of public goods can be usefully summarized into five elements: (i) information generation, (ii) infrastructure provision, (iii) political bargaining, (iv) enforcement and (v) institutional redesign. The first element, information generation, asks engineers and scientists to generate *high-quality*, *granular*, *timely*, *trustworthy* and *understandable* information on ecosystem status and change, values, uncertainty, and the magnitude and direction of human impacts.

Although there exists an example of the five elements working together to achieve single-species conservation (see the electronic supplementary material: 'Dietz *et al.*'s five elements'), to our knowledge, there is so far 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 information on many species at once are complex. This complexity is a barrier to uptake, delaying the institutional redesigns that could operationalize, finance and scale-up conservation.

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

We use the management of national forests in the United States (US) 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 use of natural resources to satisfy multiple competing interests and to maintain the natural resources in perpetuity [15–17]. 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 [9,18].

(b) 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) [5,6,9,19–24]. The eDNA methods start with DNA-based taxonomic assignment ('DNA barcoding' [25]) 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 [26,27]. 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 [5,6]. Alternatively, DNA from arthropods can be extracted *en masse* from traps [28] or from environmental substrates, such as water washes of flowers (e.g. [29]) and mass-sequenced. These latter processing pipelines are known as 'metabarcoding' or 'metagenomics', depending on whether the target DNA-barcode sequence is polymerase chain reaction-amplified (both described in [9]).

The eDNA- and sensor-based methods can all produce 'novel community data', which Hartig *et al.* [30] 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 [31,32] to presence/absence, and because the methods are automated and standardized, 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.

(c) 'Sideways' biodiversity modelling 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 series) to produce continuous spatio(-temporal) biodiversity data products [9,30,33–40]. 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 [40–43]). We were able to produce distribution maps for 76 arthropod species across a forested landscape. Because this landscape is characterized 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.* [9] and more formally described by Pollock *et al.* [44] under the name 'sideways' biodiversity modelling. In short, sideways biodiversity models (i) integrate 'the largely independent fields of biodiversity modelling and conservation' [44, p. 1119] and (ii) 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

Downloaded from https://royalsocietypublishing.org/ on 06 May 2024

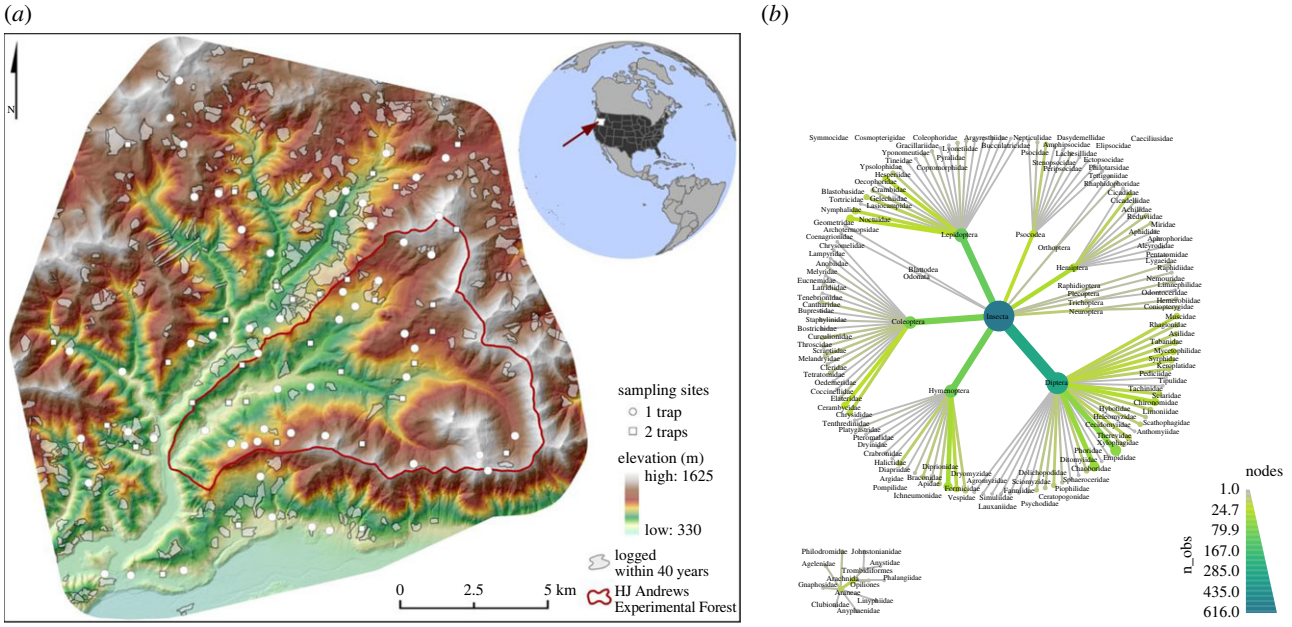


Figure 1. Sampling design and taxonomic diversity of the Malaise trapping campaign. (*a*) Sampling points in and around the H.J. Andrews Experimental Forest (red line), OR, USA. The study area consists of old-growth and logged (grey patches) deciduous and evergreen forest under different management regimes. 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-colour gradient. (*b*) Taxonomic distribution of all detected operational taxonomic units (OTUs) from the samples. Node size and colour are scaled to the number of OTUs. See the electronic supplementary material, figure S4 for a heat tree of the 190 included OTUs.

our sampling points, which creates a continuous map of arthropod biodiversity that we can use to study arthropod ecology and guide conservation.

2. Material and methods

In short, we combine DNA-based species detections, remote-sensing-derived environmental predictors, and joint species distribution modelling to predict and visualize the fine-scale distribution of arthropods across a large forested landscape. We use the joint predictions from the JSDM 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 electronic supplementary material: Materials and Methods.

(a) Model Inputs

(i) 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), OR, 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.

(ii) Wet-laboratory pipeline and bioinformatics

(iii) 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.

(iv) Creating a barcode reference database using Kelpie *in silico* polymerase chain reaction

On the output fastq files, we carried out 'in silico' PCR using `Kelpie 2.0.11` [45] and the BF3 + BR2 primers from [46], outputting 5560 unique DNA-barcode sequences. After 97%-similarity clustering and filtering for erroneous sequences, we were left with 1225 operational taxonomic units (OTUs) as the reference barcode set.

(v) Read mapping to reference barcodes

We then mapped the 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 reads mapped at high quality along more than 50% of its barcode length, following acceptance criteria from Ji *et al.* [47].

Downloaded from https://royalsocietypublishing.org/ on 06 May 2024

(vi) Environmental covariates

To predict species occurrences in the areas between the sampling points, we collected 58 continuous-space predictors (electronic supplementary material, table S1), 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 [48]. 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. 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 [49]. The topography variables were calculated from LiDAR ground returns, including elevation, slope, eastness and northness split from aspect, topographic position index, topographic roughness index (TRI) [50], topographic wetness index [51] and distance to streams, based on a vector stream network (http://oregonexplorer.info, accessed 24 October 2019). The anthropogenic variables include distance to nearest road, proportion of area logged within the last 100 and within the last 40 years, within radii of 250, 500 and 1000 m, and a categorical variable of inside or outside the boundary of the HJA. 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 (electronic supplementary material, table S1 and figure S5), which we mapped across the study area at 30 m resolution.

(b) Statistical analyses

(i) Species inputs

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

(ii) Joint species distribution model

The general idea behind species distribution modelling is to 'predict a species' distribution'. We use each species' observed incidences (1/0) at all sampling points, plus the environmental-covariate values at those points, to 'fit' a model that predicts the species' incidences from the covariate values. Once we have a fitted model, we use it to predict the species' probability of presence over the rest of the sampling area, where the environmental-covariate values are known but the species' incidences are not. Spatial autocorrelation was accounted by a trend-surface component. JSDMs extend individual species distribution models by additionally accounting for co-occurrences of species (see the electronic supplementary material: Joint Species Distribution Model).

(iii) 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 minimize overfitting, we used regularization and cross-validation. Regularization 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 (tuning). 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 ($n = 30$) (electronic supplementary material, figure S1). The training data were used to try 1000 different hyperparameter combinations in a fivefold cross-validation design, some of which are the penalty terms, to find the combination that achieves the highest predictive performance on the training data itself (see the electronic supplementary material: Tuning and Testing, figure S1). 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 JSDM `R` package `sjSDM 1.0.5` [42], with the DNN deep neural network (DNN) option to account for complex, nonlinear effects of environmental covariates (the DNN outperformed a linear model; see the electronic supplementary material, figure S11), which suits our dataset of many species with few data points and many covariates.

Finally, to estimate how OTU incidence affects the variability of predictive accuracies, we also tuned a model to the whole dataset in a fivefold cross-validation, found optimal hyperparameters, and used them in another fivefold cross-validation on the entire dataset to estimate the variability of predictive area under the curve (AUCs) by OTU (see the electronic supplementary material: Variability in Predictive AUC by OTU Incidence). We emphasize that method is only useful for estimating variability in predictive performance, given that it potentially overestimates predictive performance, which is what we avoided by using a pure holdout in the main analysis.

(iv) Variable importance with explainable-artificial intelligence

The mathematical functions used in neural network 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-artificial intelligence' (xAI) analysis, using the `R` package `flashlight 0.8.0` [52]. 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 (see the electronic supplementary material: Variable importance with explainable AI (xAI)).

(v) Prediction and visualization 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 of this one model is 76 individual and

Downloaded from https://royalsocietypublishing.org/ on 06 May 2024

continuous species distribution maps, which we combined 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-distributed stochastic neighbour embedding (T-SNE) method [53,54] to summarize species compositional change across the landscape. Pixels that have similar species compositions receive similar T-SNE values, which can be visualized. Third, we calculated Baisero *et al.*'s [55] 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 [56] and inside/outside HJA.

3. Results

(a) Model inputs

(i) DNA/taxonomic data

The 121 samples from July 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 the 190 OTUs used in our JSDM, 183 were assigned to Insecta, and seven to Arachnida (figure 1*b*). All OTUs could be assigned to order level, 178 to family level, 131 to genus level and 66 to species level (figure 1*b*; electronic supplementary material, figure S4).

(b) Statistical analyses

(i) Model performance and xAI

Across all species together, the final JSDM model achieves median and mean explanatory-performance values of AUC = 0.86 and 0.86, respectively, where the AUC 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 (electronic supplementary material, figure S2*a*). 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 species' taxonomic family or prevalence (per cent presence in sampling points).

Mean predictive AUC value does not increase with OTU abundance (as measured by incidence), and variability in predictive AUC values is only weakly higher in low-incidence OTUs (electronic supplementary material, figure S12), especially for the OTUs with high mean predictive AUCs (i.e. those used to map species richness, composition and site irreplaceability).

Out of 29 environmental covariates, 18 (electronic supplementary material, table S1) were the most important for at least one species (electronic supplementary material, figure S2*b*). Elevation and 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 (electronic supplementary material, figure S8).

(ii) 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 (figure 2*e–l*; electronic supplementary material, figure S9). We then combined 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' [57]. 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 HJA, on west-facing slopes, and in the south of the study area (figure 2*a*). A *post hoc* analysis found a nonlinear 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 1*a* and 3*c*), whereas T-SNE-2 (like species richness) appears to be correlated with the extent of surrounding old-growth forest, but only 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*).

4. 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 (electronic supplementary material, figure S5) from Landsat, LiDAR and other layers that covered information on forest structure, vegetation condition, topography and anthropogenic impact. We used a JSDM with a DNN to predict the fine-scale spatial distributions of 76 Insecta and Arachnida species with a

Downloaded from https://royalsocietypublishing.org/ on 06 May 2024

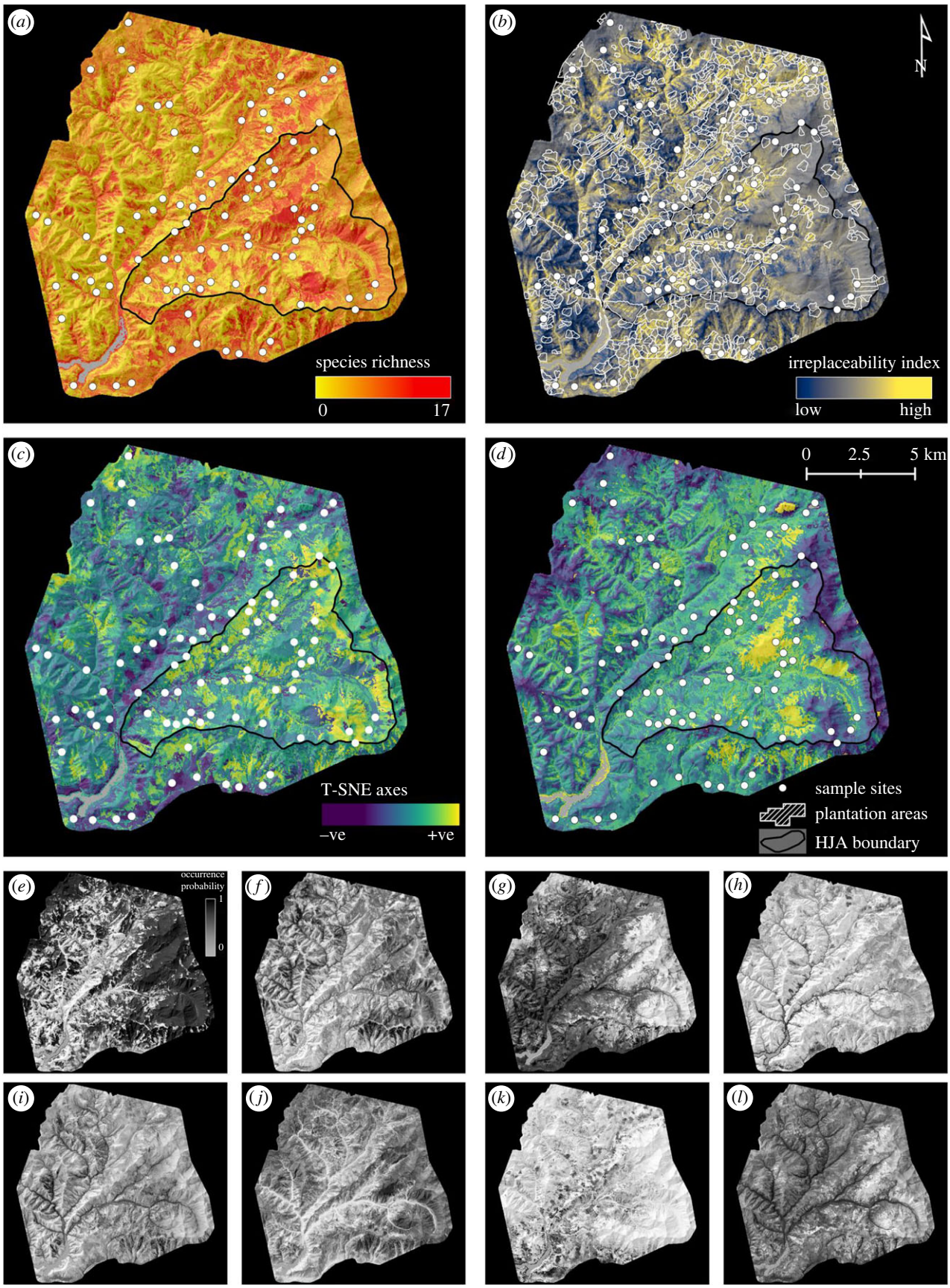


Figure 2. JSDM-interpolated spatial variation in species richness, irreplaceability, and composition, plus examples of individual species distributions. (*a*) Species richness. (*b*) Site beta irreplaceability, showing areas of forest plantation. (*c,d*) T-SNE axes 1 and 2. White circles indicate sampling points, white polygons indicate plantation areas (i.e. a record of logging in the last 100 years), and the black-line-bordered triangular area delimits the H.J. Andrews Experimental Forest (HJA; figure 1). (*e–l*) Selected individual species distributions (all species in the electronic supplementary material, figure S9), 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).

Downloaded from https://royalsocietypublishing.org/ on 06 May 2024

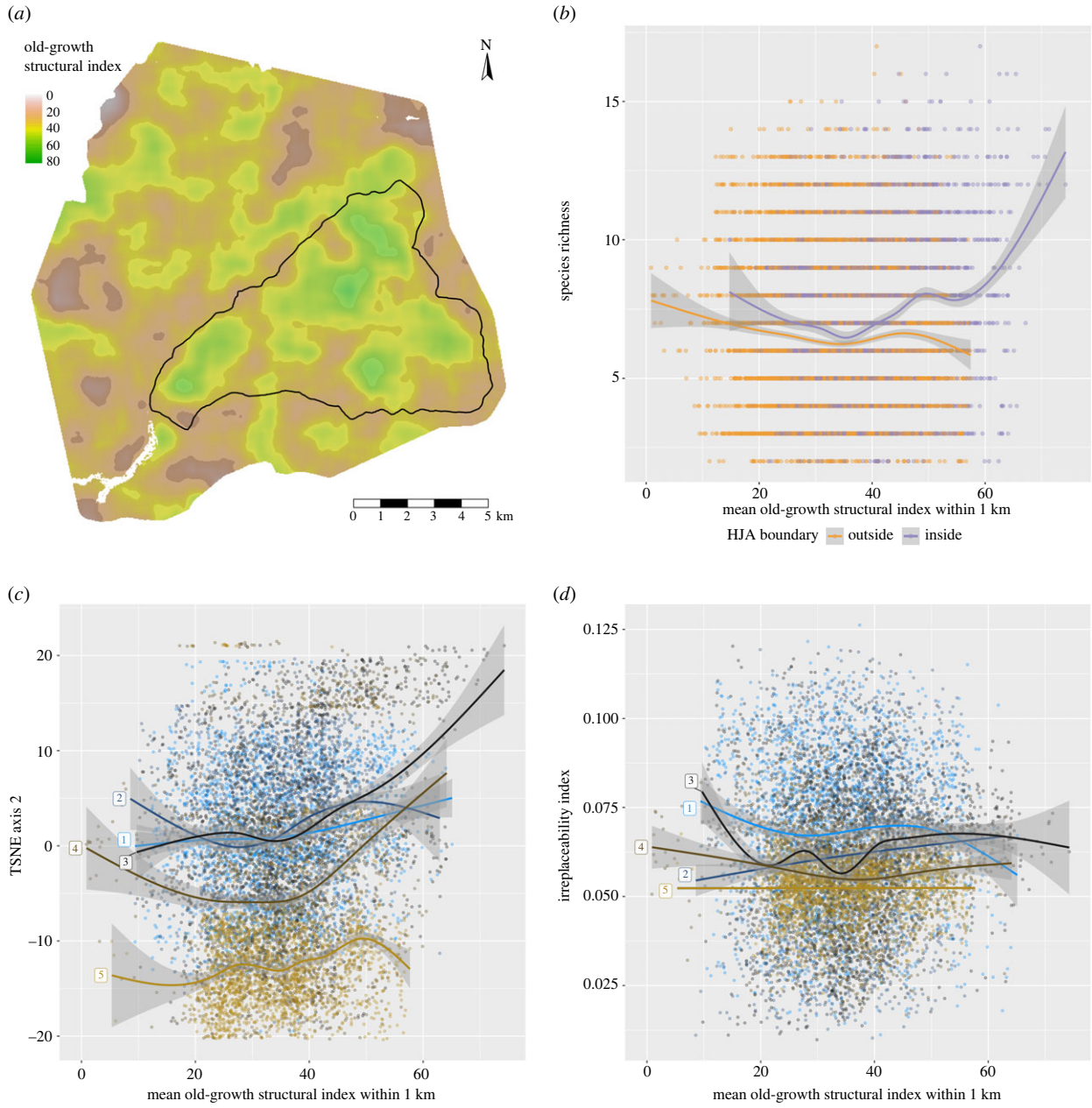


Figure 3. *Post hoc* analysis of species richness, composition and irreplaceability patterns in figure 2, in relation to an old-growth structural index (OGSI) map, from Davis *et al.* [56]. (*a*) Smoothed OGSI, showing principal patches of old-growth forest inside and outside the H.J. Andrews Experimental Forest (HJA; black-line-bordered triangular area). 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` [58].

high degree of estimated predictive performance (all individual predictive AUCs > 0.7, mean = 0.834; electronic supplementary material, figure S2*a*). The model made good use of the 29 environmental covariates, with 18 of them being the most important for at least one species (electronic supplementary material, figure S2*b*), with elevation and 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 (electronic supplementary material, figure S8).

By interpolating to create continuous species distribution maps and combining them, we created *granular* maps of arthropod biodiversity metrics: species richness, community composition and site irreplaceability (figure 2). 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 [55] 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 (electronic supplementary material, figure S9). 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

Downloaded from https://royalsocietypublishing.org/ on 06 May 2024

commercial laboratories in some countries with turnaround times measured in weeks. Information *quality* can be assessed via prediction performance (electronic supplementary material, figure S2*a*), 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 [30].

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 standardized enough to scale up to thousands of square kilometres. However, as shown by the many species with low predictive AUCs (electronic supplementary material, figure S2*a*), future work will be needed to improve how error is accounted for when generating model outputs [30,32], and we discuss methods for doing this in the electronic supplementary material: Caveats. We conclude by briefly reviewing potential applications of this approach.

(a) 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 [9,19,24,59]. 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 [60], and graphics processing unit-accelerated deep learning algorithms can be used to both accelerate and improve model fitting on these larger datasets [42,61]. 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 [20,29,36,62–68], with river networks being an especially promising way to scale up sampling over large areas [63,69].

As a result, it is possible to envisage implementing Pollock *et al.*'s [44] vision of using 'sideways' species-based biodiversity monitoring to subdivide whole landscapes for ranking by conservation value (see also [38]). 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 [70].

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.* [64] 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 [71]. Chiaverini *et al.* [72] 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 the International Union for Conservation of Nature range maps, which are based on historical distribution data (https://www.iucnredlist.org, accessed 18 April 2022). Finally, Hamilton *et al.* [3] compiled decades of standardized biodiversity inventory data for 2216 species in the continental USA and interpolated to identify areas of unprotected biodiversity importance (using a measure similar to site irreplaceability, i.e. protection-weighted range-size rarity). Because the resulting maps were *granular* (990 m), Hamilton *et al.* [3] 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 owing to the inclusion of taxa not normally included in such analyses (especially plants, freshwater invertebrates and pollinators).

(b) Conclusion

A major difficulty for basic and applied community ecology is the collection of many standardized observations of many species. DNA-based methods provide capacity for collecting data on many species at once, but costs scale with sample number. By 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.

Data accessibility. Raw sequence data are archived at NCBI Short Read Archive BioProject PRJNA869351. All scripts and data tables (from bioinformatic processing to statistical analysis to figure generation) are available from the GitHub respository: https://github.com/chnpenny/HJA_analyses_Kelpie_clean/releases/tag/v1.1.0 and archived at https://zenodo.org/records/8303158 [73].

Supplementary material is available online [74].

Declaration of AI use. Yes, we have used AI-assisted technologies in creating this article.

Authors' contributions. Y.L.: conceptualization, formal analysis, investigation, methodology, validation, visualization, writing—original draft, writing—review and editing; C.D.: conceptualization, formal analysis, investigation, methodology, validation, visualization, writing—original draft, writing—review and editing; M.I.T.: conceptualization, data curation, investigation, methodology, project administration, writing—review and editing; M.L.: investigation, methodology, writing—review and editing; D.M.B.: project administration, resources, supervision, writing—review and editing; D.B.L.: project administration, resources, supervision, writing—review and editing; P.G.: software; M.P.: methodology, software, validation, visualization, writing—review and editing; T.L.: conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, visualization, writing—review and editing; D.W.Y.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. D.W.Y. is a co-founder of NatureMetrics (www.naturemetrics.com), which provides commercial metabarcoding services. All other authors have no competing interests.

Funding. D.W.Y. and M.L. 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

Downloaded from https://royalsocietypublishing.org/ on 06 May 2024

(GREKF19-01, GREKF20-01 and GREKF21-01) at the Kunming Institute of Zoology, the Yunnan Revitalization Talent Support Program: High-end Foreign Expert Project, and the University of Chinese Academy of Sciences. D.W.Y. 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. M.I.T. was supported by the National Science Foundation-funded H.J. Andrews Long-Term Ecological Research (LTER) program (no. DEB-1440409), Oregon State University, the ARCS Oregon Chapter and the US Department of Agriculture Forest Service. Field data collection was funded by Oregon State University, the Pacific Northwest Research Station and the US Department of Agriculture Forest Service. LiDAR data processing was supported by the National Science Foundation-funded H.J. Andrews LTER program (nos. 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 US Department of Agriculture or US 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 US Government of any product or service.

Acknowledgements. We thank field technicians B. P. Murley, S. D. Sparrow and M. E. Yates.

References

1. Prather CM *et al.* 2013 Invertebrates, ecosystem services and climate change: invertebrates, ecosystems and climate change. *Biol. Rev.* **88**, 327–348. (doi:10.1111/brv.12002)
2. Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. 2017 Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **7**, 9132. (doi:10.1038/s41598-017-09084-6)
3. Hamilton H *et al.* 2022 Increasing taxonomic diversity and spatial resolution clarifies opportunities for protecting US imperiled species. *Ecol. Appl.* **32**, e2534. (doi:10.1002/eap.2534)
4. Westgate MJ, Barton PS, Lane PW, Lindenmayer DB. 2014 Global meta-analysis reveals low consistency of biodiversity congruence relationships. *Nat. Commun.* **5**, 3899. (doi:10.1038/ncomms4899)
5. Chua PY, Bourlat SJ, Ferguson C, Korlevic P, Zhao L, Ekrem T, Meier R, Lawniczak MK. 2023 Future of DNA-based insect monitoring. *Trends Genet.* **39**, 531–544. (doi:10.1016/j.tig.2023.02.012)
6. van Klink R *et al.* 2022 Emerging technologies revolutionise insect ecology and monitoring. *Trends Ecol. Evol.* **37**, 872–885. (doi:10.1016/j.tree.2022.06.001)
7. Lewinsohn TM, Roslin T. 2008 Four ways towards tropical herbivore megadiversity. *Ecol. Lett.* **11**, 398–416. (doi:10.1111/j.1461-0248.2008.01155.x)
8. Zhang K *et al.* 2016 Plant diversity accurately predicts insect diversity in two tropical landscapes. *Mol. Ecol.* **25**, 4407–4419. (doi:10.1111/mec.13770)
9. Bush A *et al.* 2017 Connecting Earth observation to high-throughput biodiversity data. *Nat. Ecol. Evol.* **1**, 0176. (doi:10.1038/s41559-017-0176)
10. Bae S *et al.* 2019 Radar vision in the mapping of forest biodiversity from space. *Nat. Commun.* **10**, 4757. (doi:10.1038/s41467-019-12737-x)
11. Müller J, Moning C, Bässler C, Heurich M, Brandl R. 2009 Using airborne laser scanning to model potential abundance and assemblages of forest passerines. *Basic Appl. Ecol.* **10**, 671–681. (doi:10.1016/j.baae.2009.03.004)
12. Müller J, Brandl R. 2009 Assessing biodiversity by remote sensing in mountainous terrain: the potential of LiDAR to predict forest beetle assemblages. *J. Appl. Ecol.* **46**, 897–905. (doi:10.1111/j.1365-2664.2009.01677.x)
13. Rhodes MW, Bennie JJ, Spalding A, Maclean IMD. 2022 Recent advances in the remote sensing of insects. *Biol. Rev.* **97**, 343–360. (doi:10.1111/brv.12802)
14. Dietz T, Ostrom E, Stern PC. 2003 The struggle to govern the commons. *Science* **302**, 1907–1912. (doi:10.1126/science.1091015)
15. Carter SK *et al.* 2019 Quantifying ecological integrity of terrestrial systems to inform management of multiple-use public lands in the United States. *Environ. Manage.* **64**, 1–19. (doi:10.1007/s00267-019-01163-w)
16. Hobbs RJ *et al.* 2010 Guiding concepts for park and wilderness stewardship in an era of global environmental change. *Front. Ecol. Environ.* **8**, 483–490. (doi:10.1890/090089)
17. Loomis J. 2002 *Integrated public lands management: principles and applications to national forests, parks, wildlife refuges, and BLM lands*. New York, NY: Columbia University Press.
18. Frankham R. 2010 Challenges and opportunities of genetic approaches to biological conservation. *Biol. Conserv.* **143**, 1919–1927. (doi:10.1016/j.biocon.2010.05.011)
19. Besson M, Alison J, Bjerge K, Gorochowski TE, Høye TT, Jucker T, Mann HMR, Clements CF. 2022 Towards the fully automated monitoring of ecological communities. *Ecol. Lett.* **25**, 2753–2775. (doi:10.1111/ele.14123)
20. Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, Yu DW, de Bruyn M. 2014 Environmental DNA for wildlife biology and biodiversity monitoring. *Trends Ecol. Evol.* **29**, 358–367. (doi:10.1016/j.tree.2014.04.003)
21. Christin S, Hervet E, Lecomte N. 2019 Applications for deep learning in ecology. *Methods Ecol. Evol.* **10**, 1632–1644. (doi:10.1111/2041-210X.13256)
22. Pawlowski J, Apothéloz-Perret-Gentil L, Altermatt F. 2020 Environmental DNA: what's behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Mol. Ecol.* **29**, 4258–4264. (doi:10.1111/mec.15643)
23. Ruppert KM, Kline RJ, Rahman MS. 2019 Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring, and applications of global eDNA. *Global Ecol. Conserv.* **17**, e00547. (doi:10.1016/j.gecco.2019.e00547)
24. Tosa MI *et al.* 2021 The rapid rise of next-generation natural history. *Front. Ecol. Evol.* **9**, 698131. (doi:10.3389/fevo.2021.698131)
25. Hebert PDN, Cywinska A, Ball SL. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B* **270**, 313–321. (doi:10.1098/rspb.2002.2218)
26. Ratnasingham S. 2019 mBRAVE: the multiplex barcode research and visualization environment. *Biodivers. Inf. Sci. Stand.* **3**, e37986. (doi:10.3897/biss.3.37986)
27. Srivathsan A, Lee L, Katoh K, Hartop E, Kutty SN, Wong J, Yeo D, Meier R. 2021 ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. *BMC Biol.* **19**, 217. (doi:10.1186/s12915-021-01141-x)
28. Ji Y *et al.* 2013 Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.* **16**, 1245–1257. (doi:10.1111/ele.12162)
29. Thomsen PF, Sigsgaard EE. 2019 Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. *Ecol. Evol.* **9**, 1665–1679. (doi:10.1002/ece3.4809)
30. Hartig F *et al.* 2024 Novel community data–properties and prospects. *Trends Ecol. Evol.* **39**, 280–293. (doi:10.1016/j.tree.2023.09.017)
31. Luo M, Ji Y, Warton D, Yu DW. 2023 Extracting abundance information from DNA-based data. *Mol. Ecol. Resour.* **23**, 174–189. (doi:10.1111/1755-0998.13703)
32. Diana A, Matechou E, Griffin J, Yu DW, Luo M, Tosa M, Bush A, Griffiths R. 2022 eDNAPlus: a unifying modelling framework for DNA-based biodiversity monitoring. (http://arxiv.org/abs/2211.12213 [stat])
33. He KS, Bradley BA, Cord AF, Rocchini D, Tuanmu M, Schmidtlein S, Turner W, Wegmann M, Pettorelli N. 2015 Will remote sensing shape the next generation of species distribution models? *Remote Sens. Ecol. Conserv.* **1**, 4–18. (doi:10.1002/rse2.7)
34. Kwok R. 2018 Ecology's remote-sensing revolution. *Nature* **556**, 137–138. (doi:10.1038/d41586-018-03924-9)
35. Leitão PJ, Santos MJ. 2019 Improving models of species ecological niches: a remote sensing overview. *Front. Ecol. Evol.* **7**, 9. (doi:10.3389/fevo.2019.00009)
36. Lin M *et al.* 2021 Landscape analyses using eDNA metabarcoding and Earth observation predict community biodiversity in California. *Ecol. Appl.* **31**, e02379. (doi:10.1002/eap.2379)

Downloaded from https://royalsocietypublishing.org/ on 06 May 2024

37. Pettorelli N *et al.* 2018 Satellite remote sensing of ecosystem functions: opportunities, challenges and way forward. *Remote Sens. Ecol. Conserv.* **4**, 71–93. (doi:10.1002/rse2.59)
38. Cavender-Bares J *et al.* 2022 Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nat. Ecol. Evol.* **6**, 506–519. (doi:10.1038/s41559-022-01702-5)
39. Müller J *et al.* 2023 Soundscapes and deep learning enable tracking biodiversity recovery in tropical forests. *Nat. Commun.* **14**, 6191. (doi:10.1038/s41467-023-41693-w)
40. Davis CL, Bai Y, Chen D, Robinson O, Ruiz-Gutierrez V, Gomes CP, Fink D. 2023 Deep learning with citizen science data enables estimation of species diversity and composition at continental extents. *Ecology* **104**, e4175. (doi:10.1002/ecy.4175)
41. Ovaskainen O, Abrego N. 2020 *Joint species distribution modelling: with applications in R*, 1st edn. Cambridge, UK: Cambridge University Press.
42. Pichler M, Hartig F. 2021 A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods Ecol. Evol.* **12**, 2159–2173. (doi:10.1111/2041-210X.13687)
43. Warton DI, Blanchet FG, O'Hara RB, Ovaskainen O, Taskinen S, Walker SC, Hui FK. 2015 So many variables: joint modeling in community ecology. *Trends Ecol. Evol.* **30**, 766–779. (doi:10.1016/j.tree.2015.09.007)
44. Pollock LJ, O'Connor LM, Mokany K, Rosauer DF, Talluto MV, Thuiller W. 2020 Protecting biodiversity (in all its complexity): new models and methods. *Trends Ecol. Evol.* **35**, 1119–1128. (doi:10.1016/j.tree.2020.08.015)
45. Greenfield P, Tran-Dinh N, Midgley D. 2019 Kelpie: generating full-length 'amplicons' from whole-metagenome datasets. *PeerJ* **6**, e6174. (doi:10.7717/peerj.6174)
46. Elbrecht V, Braukmann TW, Ivanova NV, Prosser SW, Hajibabaei M, Wright M, Zakharov EV, Hebert PD, Steinke D. 2019 Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ* **7**, e7745. (doi:10.7717/peerj.7745)
47. Ji Y, Huotari T, Roslin T, Schmidt NM, Wang J, Yu DW, Ovaskainen O. 2020 SPIKEPIPE: a metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Mol. Ecol. Resour.* **20**, 256–267. (doi:10.1111/1755-0998.13057)
48. Kane VR, McGaughey RJ, Bakker JD, Gersonde RF, Lutz JA, Franklin JF. 2010 Comparisons between field- and LiDAR-based measures of stand structural complexity. *Can. J. For. Res.* **40**, 761–773. (doi:10.1139/X10-024)
49. Müller J, Bae S, Röder J, Chao A, Didham RK. 2014 Airborne LiDAR reveals context dependence in the effects of canopy architecture on arthropod diversity. *Forest Ecol. Manage.* **312**, 129–137. (doi:10.1016/j.foreco.2013.10.014)
50. Wilson MFJ, O'Connell B, Brown C, Guinan JC, Grehan AJ. 2007 Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. *Mar. Geodesy* **30**, 3–35. (doi:10.1080/01490410701295962)
51. Metcalfe P, Beven K, Freer J. 2018. dynatopmodel: implementation of the dynamic TOPMODEL hydrological model. See https://cran.r-project.org/src/contrib/Archive/dynatopmodel/dynatopmodel_1.2.1.tar.gz.
52. Mayer M. 2021 flashlight: shed light on black box machine learning models. R package version 0.8.0. See https://cran.r-project.org/package=flashlight.
53. van der Maaten L, Hinton G. 2008 Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605.
54. Krijthe JH. 2015 Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. R package version 0.15. See https://cran.r-project.org/web/packages/Rtsne/Rtsne.pdf.
55. Baisero D, Schuster R, Plumptre AJ. 2022 Redefining and mapping global irreplaceability. *Conserv. Biol.* **36**, e13806. (doi:10.1111/cobi.13806)
56. Davis RJ, Ohmann JL, Kennedy RE, Cohen WB, Gregory MJ, Yang Z, Roberts HM, Gray AN, Spies TA. 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, USA.
57. Kukkala AS, Moilanen A. 2013 Core concepts of spatial prioritisation in systematic conservation planning. *Biol. Rev.* **88**, 443–464. (doi:10.1111/brv.12008)
58. Wood S. 2017 *Generalized additive models: an introduction with R*, 2nd edn. New York, NY: Chapman and Hall/CRC.
59. Speaker T *et al.* 2022 A global community-sourced assessment of the state of conservation technology. *Conserv. Biol.* **36**, e13871. (doi:10.1111/cobi.13871)
60. Lock M, van Duren I, Skidmore AK, Saintilan N. 2022 Harmonizing forest conservation policies with essential biodiversity variables incorporating remote sensing and environmental DNA technologies. *Forests* **13**, 445. (doi:10.3390/f13030445)
61. Pichler M, Hartig F. 2023 Machine learning and deep learning: a review for ecologists. *Methods Ecol. Evol.* **14**, 994–1016. (doi:10.1111/2041-210X.14061)
62. Abrego N, Norros V, Halme P, Somervuo P, Ali-Kovero H, 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. *Mol. Ecol. Resour.* **18**, 511–524. (doi:10.1111/1755-0998.12755)
63. Guimaræes Sales N *et al.* 2020 Fishing for mammals: landscape-level monitoring of terrestrial and semi-aquatic communities using eDNA from riverine systems. *J. Appl. Ecol.* **57**, 707–716. (doi:10.1111/1365-2664.13592)
64. Ji Y *et al.* 2022 Measuring protected-area effectiveness using vertebrate distributions from leech iDNA. *Nat. Commun.* **13**, 1555. (doi:10.1038/s41467-022-28778-8)
65. Leempoel K, Hebert T, Hadly EA. 2019 A comparison of eDNA to camera trapping for assessment of terrestrial mammal diversity. *Ecology* (preprint). (doi:10.1101/634022).
66. Massey AL *et al.* 2022 Invertebrates for vertebrate biodiversity monitoring: comparisons using three insect taxa as iDNA samplers. *Mol. Ecol. Resour.* **22**, 962–977. (doi:10.1111/1755-0998.13525)
67. Rodgers TW *et al.* 2017 Carrion fly-derived DNA metabarcoding is an effective tool for mammal surveys: evidence from a known tropical mammal community. *Mol. Ecol. Resour.* **17**, e133–e145. (doi:10.1111/1755-0998.12701)
68. Tilker A *et al.* 2020 Identifying conservation priorities in a defaunated tropical biodiversity hotspot. *Divers. Distrib.* **26**, 426–440. (doi:10.1111/ddi.13029)
69. Lyet A, Pellissier L, Valentini A, Dejean T, Hehmeyer A, Naidoo R. 2021 eDNA sampled from stream networks correlates with camera trap detection rates of terrestrial mammals. *Sci. Rep.* **11**, 11362. (doi:10.1038/s41598-021-90598-5)
70. Weston P. 2021 New biodiversity algorithm 'will blight range of natural habitats in England'. *The Guardian*. See https://www.theguardian.com/environment/2021/jul/21/biodiversity-metric-algorithm-natural-england-developers-blight-valuable-habitats-aoe.
71. Maxwell SL *et al.* 2020 Area-based conservation in the twenty-first century. *Nature* **586**, 217–227. (doi:10.1038/s41586-020-2773-z)
72. Chiaverini L *et al.* 2022 Multi-scale, multivariate community models improve designation of biodiversity hotspots in the Sunda Islands. *Anim. Conserv.* **25**, acv.12771. (doi:10.1111/acv.12771)
73. Li Y *et al.* 2024 Data from: Combining environmental DNA and remote sensing for efficient, fine-scale mapping of arthropod biodiversity. *Zenodo* (https://zenodo.org/records/8303158)
74. Li Y *et al*. 2024 Combining environmental DNA and remote sensing for efficient, fine-scale mapping of arthropod biodiversity. Figshare. (doi:10.6084/m9.figshare.c.7151335)

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

in Philosophical Transactions of the Royal Society B

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 et al.'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 1a). 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

Before extraction, we kept only the heads of insects with body sizes longer than 2 cm. 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 (`FilterReads -qt 30 +f GenBank_24919_COI_C99_20.mer 25pct input.fq`), using a reference kmer dataset `GenBank_24919_COI_C99_20.mer` (accessed 3 Aug 2021). This step is optional but greatly increases efficiency. We then used `Kelpie 2.0.11` [Greenfield et al., 2019] to carry out *in-silico* PCR on the filtered datasets (`Kelpie -f CCHGAYATRGCHTTYCCHCG -r TCDGGRTGNCCRAARAAYCA -primers -filtered -min 400 -max 500`). Binaries for both are at `https://github.`

com/PaulGreenfieldOz/WorkingDogs/tree/master/Kelpie_v2 (accessed 20 Nov 2023). 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. The advantages are that it is trivial to switch primers, lab workload is reduced, there can be no PCR error or PCR 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 (ID'd 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 ~ 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. 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^{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 1S) 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 1b).

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 1.0.5` [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 binomial 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:

$Z_{ij} = \beta_{0j} + DNN(X_{in}) + U_{E_i}\beta_{E_j} + U_{N_i}\beta_{N_j} + (U_{E_i}U_{N_i})\beta_{EN_j} + MVN(0, \Sigma_{ij})$

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

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. The second part of the model describes the trend-surface model, which is one way to account for spatial auto-correlation [Dormann et al., 2007]: U_{E_i} and U_{N_i} are the two Universal Transverse Mercator variables (coordinates) which are modeled for each species j at sampling site i as linear terms with coefficients β_{E_j} and β_{N_j}, and as interaction with coefficients β_{EN_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, 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.

To minimise the risk of overfitting, we applied a combination of regularisation and cross validation. Regularisation is a statistical method that reduces small (or uncertain or collinear) covariate effects to zero. In this way, the initially high complexity of a DNN algorithm can end in a DNN model with a low effective complexity with good generality, even for small data.

In Figure $1SB_2$, we list nine model 'hyperparameters', which consist of the weighting between lasso and ridge regularisation ($\alpha_{e,s,b}$) and their strengths ($\lambda_{e,s,b}$) for each of the environmental, spatial, and species covariance components, plus the dropout rate, the hidden structure for the DNN, and the learning rate of the model (Figure

1SC). These hyperparameters govern the neural network's structure and how it is fit to the data, and the challenge in fitting is to select optimal regularisation values (the alphas and lambdas, and the dropout rate) for accurate prediction, which we do via 5-fold, nested cross-validation, in a procedure known as model tuning.

First, we randomly split the 121 data points from July 2018 into 75% *training* data ($n = 91$) and 25% *test* data ($n = 30$) (the latter also known as hold-out data, or outer split), and we ensured that when two Malaise traps had been placed at the same site, they were assigned to the same split (Figure 1SA).

We then worked with only the *training* dataset for model tuning (i.e. inner split). We split the training dataset into five 'folds' (=sections), also ensuring that data from pairs of traps placed at the same site were assigned to the same fold. We chose one combination of hyperparameter values, fit the model with those hyperparameter values to 4 of the 5 folds (as a single dataset), and measured how well this fitted model predicted presences and absences in the sites from the fifth fold (the *validation* dataset), which the model had *not* been fit to. This is the model's predictive performance on that fold with that hyperparameter combination. Because we chose **5**-fold CV, we repeated this procedure five times, each time predicting a different fold of the five (Figure $1SB_1$). We calculated the model's mean predictive performance over the five validation datasets and the model's mean explanatory performance on the five training datasets. We repeated this five-fold CV procedure for 1000 hyperparameter combinations sampled from the total set of possible hyperparameter combinations ($n = 7200$), recording all 1000 mean performances in Figure (1SC) (black pts: mean predictive performances. blue pts: mean explanatory performances). We used six metrics to evaluate predictive and explanatory performance: AUC (area under the receiver operating characteristic curve), positive likelihood ratio, Pearson's correlation coefficient, log-likelihood, True Skill Statistic (TSS), and Nagelkerke's R^2 [Lawson et al., 2014, Wilkinson et al., 2021, see Supplementary Information].

From the set of 1000 models, we chose the model with the hyperparameter combination that produced the highest predictive performance (designated as the tuned model) and fit it to the full training dataset (i.e. no folds). This is the *Final fit* model (Figure $1SB_1$), which we used to calculate explanatory performances per species. Finally, we also used *Final fit* model to predict presences/absences in the 25% *test* dataset that the model had never seen and calculated a predictive performance per species: $\mathrm{AUC_{pred}}$. The final models chosen by the other performance metrics behaved similarly (Figure 3S).

For species mapping, we filtered to those species that showed moderate to good predictive performance ($AUC_{pred} > 0.70$, $mean = 0.83$). **The key point is that because $\mathbf{AUC_{pred}}$ is calculated from the test dataset, *which the model never saw during tuning and final fitting*, we can use each species' $\mathbf{AUC_{pred}}$ as a measure of model generality for that species, and high $\mathbf{AUC_{pred}}$ species are therefore the species for which overfitting is a low risk.**

The purpose of regularisation is to create simpler, but not too simple, models, and this has the effect of creating models that are more likely to be general. When using regularisation, one is freed to use large numbers of co-

variates and terms in the model because regularisation typically sets most of their coefficients to 0. **The cost of regularisation is that one needs a large number of samples for model tuning (selecting the optimal regularisation regime via cross-validation) and to provide an untouched dataset for measuring model predictive performance.**

Variability in Predictive AUC by OTU Incidence

Finally, using a single holdout dataset for final testing does not allow estimates of the *variability* of our model (with respect to predictive AUC). We therefore ran an alternative model evaluation, using 5-fold cross validation over the whole dataset, which allows such an estimate. In this alternative evaluation, we followed the above methods to perform 5-fold cross validation, but now we used the entire dataset (121 points, 225 OTUs), with the same 75% - 25% splits for the training and validation folds, the same number of runs (1000 different combinations of hyperparameters), and the same prevalence threshold (minimum presence at 6 or more sites). From these runs, we chose the model with the highest predictive performance, as measured by AUC only in this case, and then used these hyperparameters to fit a model on the full dataset, producing the alternative *Final fit* model. Using this *Final fit* model, we ran a further 5-fold cross validation (using a 75%-25% split) and saved the results from each of the five validation predictions for all OTUs. The accuracy metrics were then averaged for each OTU and displayed graphically (Figure 12S). We ran a polynomial regression to test whether the standard deviation of predictive AUCs is greater for lower-incidence OTUs (Figure 12S). Finally, we used the OTUs with $AUC_{pred} \geq 0.70$ ($n = 112$, $AUC_{mean} = 0.80$) to create maps of species richness, ordination axes, and irreplaceability (Figure 13S), in the same way as the main analysis.

We use this alternative analysis for the *sole purpose* of estimating the *variability* of predictive AUCs because fitting a model to the whole dataset increases the risk of overfitting and could potentially overestimate the predictive performance of the model, which is what we avoided by using a pure holdout in the main analysis. Ultimately, with a much larger dataset, running a nested CV with an inner k-CV (for training) and an inner k-CV (for testing) would be the gold standard to produce reliable estimates of the predictive AUCs and their variabilities together. However, given that typical ecological community datasets have many rare and many abundant species, splitting the data twice sequentially would likely frequently produce training and test splits with either no occurrences (for rare species) or only occurrences (for abundant species), making it technically impossible to fit reliable models or to validate them fairly. This suggests that the question of how to effectively evaluate and tune ecological community models should continue to be a priority for future research.

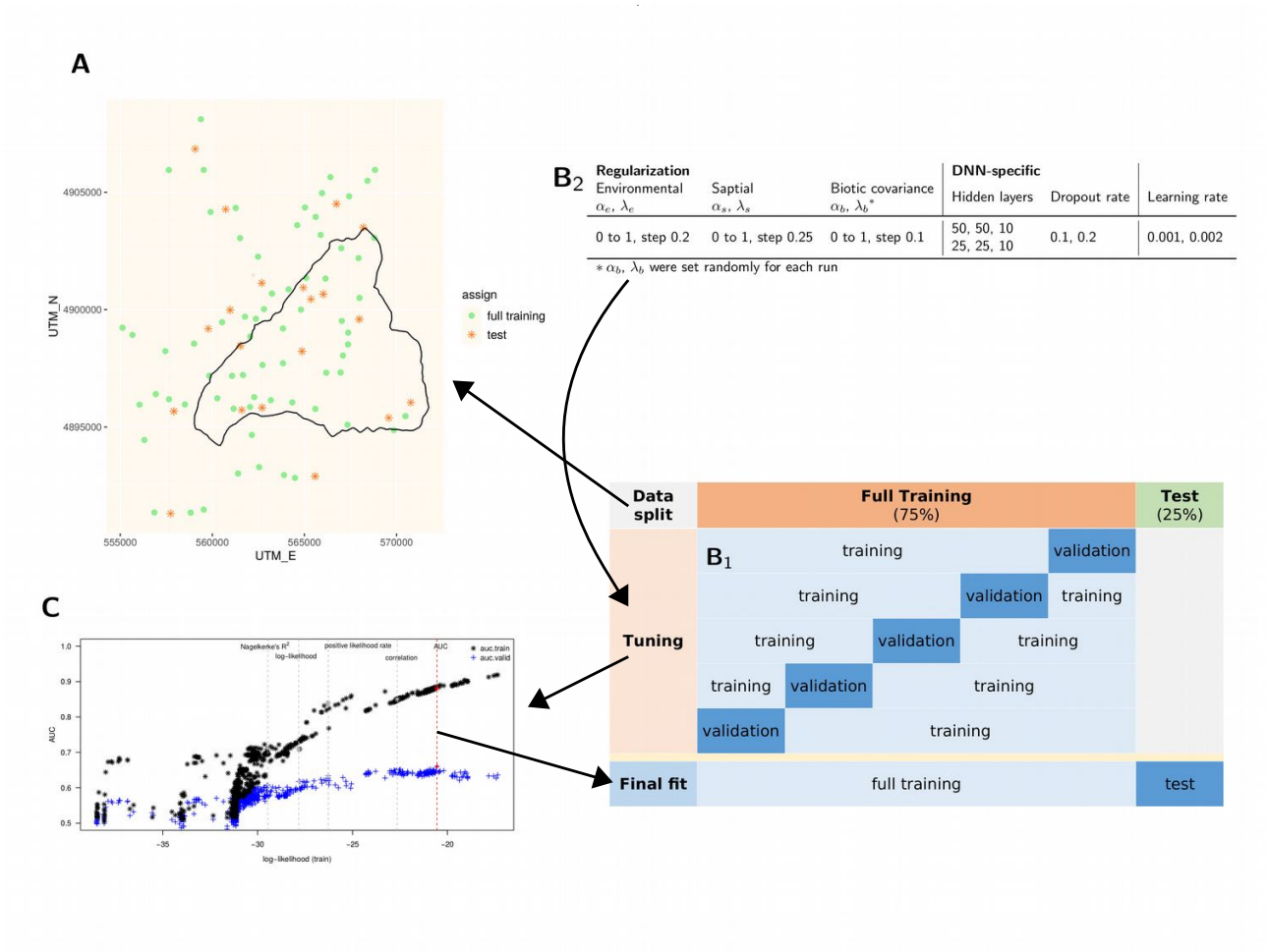


Figure 1S: Model tuning and training strategy. We obtained our final model by data splitting, tuning, and final fitting. *A*. We randomly split the 121 Malaise traps into test ($n = 30$) and training subsets (91). B_1. We then randomly split the training set into five parts for tuning via a 5-fold cross-validation. 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.

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` [Maksymiuk et al., 2020, Mayer, 2021].

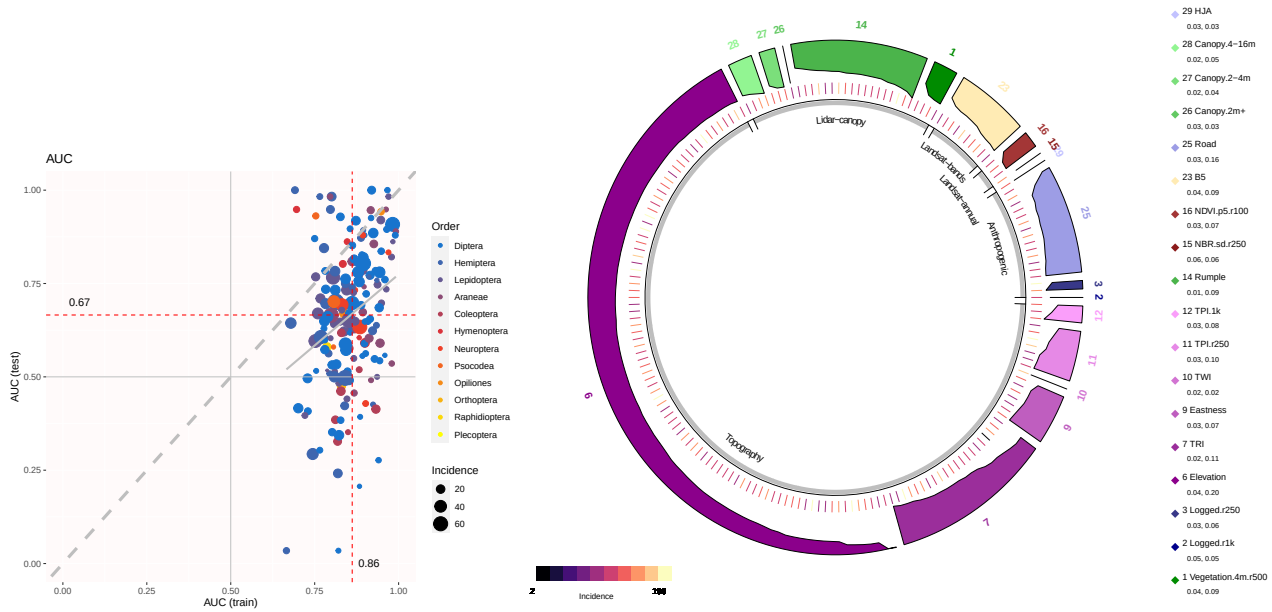
Variable importance is based on global permutations of variables in the dataset [Fisher et al., 2019]. The calculation consists of several steps: First, a variable x_i from the dataset X is permuted (the values are randomised globally (over all sites)) and replaces the original x_i in X, so that we get a new dataset $X_{permuted}$ with $x_{i,permuted}$ (all other variables are not permuted). By permuting the variable, the effect (or association) between x_i and the response variable (e.g. species occurrence) is removed. Second, we generate new predictions with our model and dataset $X_{permuted}$. Third, we calculate the predictive performance for our new predictions (here, AUC, see below). Fourth, we compare the new predictive performance for $X_{permuted}$, which contains the permuted variable $x_{i,permuted}$, with the predictive performance of the non-permuted dataset X. The difference between these two performances corresponds to the permutation importance of the variable x_i. If x_i has a strong effect on the response variable, the permutation importance of variable x_i will be large because the model cannot predict the response well anymore. All these steps are repeated for all variables in the dataset. The advantage of this variable-importance protocol is that it does not require re-fitting the model n times for the n variables in the data set. We omitted the spatial component when calculating variable importance.

Friedman's H-statistic is used to infer the importance of variable-variable interactions [Friedman and Popescu, 2008]. The statistic is based on partial dependencies (PD). Partial Dependencies describe the marginal effects of a variable on the response variable. Friedman's H statistic additively decomposes the predict function $\hat{(f)} = PD_i + PD_j + PD_{i,j}$, assuming that it consists of main effects (PD_i and PD_j) and an interaction $PD_{i,j}$ of two variables. Friedman's H statistic estimates the importance of an interaction by comparing the interaction PD with the individual PDs: $PD_{i,j} - PD_i + PD_j$. Without the subtraction, the interaction ($PD_{i,j}$) would accumulate the individual effects (we only want the "shared" part). Finally, the variance of $PD_{i,j} - PD_i + PD_j$ divided by the variance of $PD_{i,j}$ corresponds to interaction importance between x_i and x_j.

We calculated these xAI metrics based on the explanatory performance of the JSDM 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 JSDM predictions based on sjSDM [Pichler and Hartig, 2021], each species' prediction used the average of five separate prediction runs. We created binary species distributions maps 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.



(a) (b)

Figure 2S: 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.

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 combined 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 must assume that the sites selected by any given SCP run are optimal. As Langford et al. [2011] point out, SCP algorithms are not widely tested for robustness to input error.

In contrast, Baisero et al.'s (2022) site-irreplaceability value is directly calculated: defined as 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 2 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 2 A). It is important to keep in mind that any measure of site irreplaceability can only compare the sites *within* the 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 could be given low conservation targets, and artefactually rare habitats (the plantation in a rainforest) could 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 10S).

Finally, given the rapidity with which Baisero et al.'s site-irreplaceability values can be calculated, one possible approach to account for error in predicted species distributions (Figure 9S) would be to resample the site-irreplaceability calculations in some way and to plot mean or median site irreplaceability values. The idea would be to produce a map that upweights the contributions of species with higher values of predictive performance and with higher occupancy probabilities. However, this proposed approach would require testing to see whether it in fact produces a more reliable map.

False-negative and false-positive errors

Despite detecting 1225 OTUs across the whole dataset, ultimately, only 190 OTUs had $\geq$ 6 detections. 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. 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].

3. Increase the number of sampling points. This would allow the training and test dataset sizes to be increased, allow more folds in the cross-validation step, and reduce the metrics of predictive performance, since AUCpred variance decreases with incidence.

4. 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, allows one to increase the minimum incidence threshold for inclusion in the model, 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.

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]. An open question for future studies is whether it is better to include only the individual satellite spectral bands and let the DNN combine the bands, rather than also including known band combinations like NDVI.

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. The explanatory AUC of the linear model is higher than in the DNN model, but the predictive power is lower, showing more overfitting with the linear model (Figure 11S). A DNN fitting procedure thus appears to be useful for disentangling complex relationships between remote-sensing-derived environmental covariates and community data.

Going forward, new JSDM software packages are being published that can exploit sample replication to account for false negatives and false positives [Diana et al., 2022, Tobler et al., 2019, Doser et al., 2022]. Over time, as such capabilities are combined with increased efficiency, the result should be more reliable predictions.

Finally, 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.

Additional Supplementary Figures and Tables

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 ra	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`

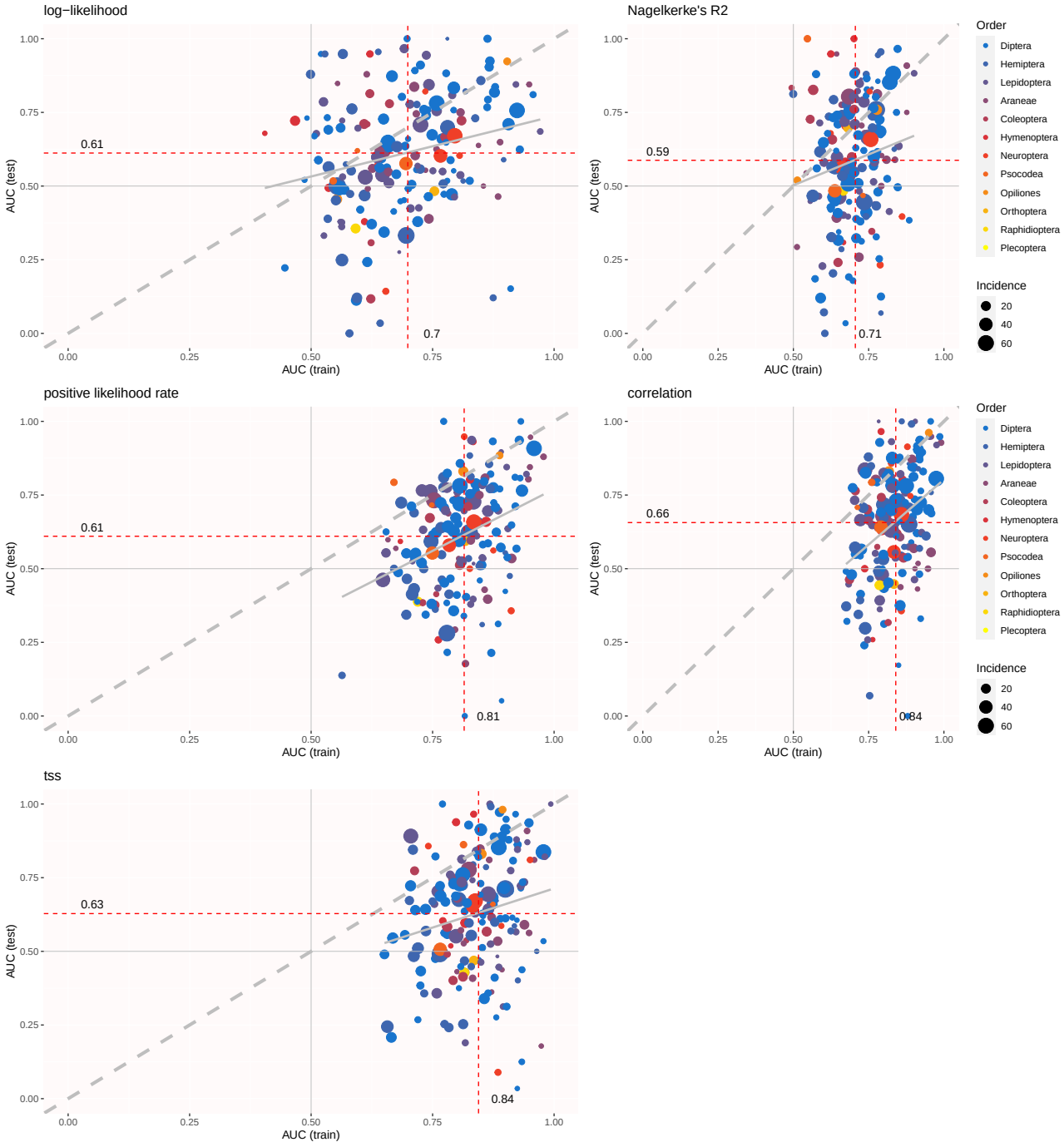


Figure 3S: Explanatory AUC vs predictive AUC for best sjSDM models tuned according to log-likelihood, Nagelkerke's R^2, positive likelihood rate, correlation and TSS(true skill statistic). Each point is one OTU. Color indicates taxonomic class (order), and point size indicates incidence (number of Malaise 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.

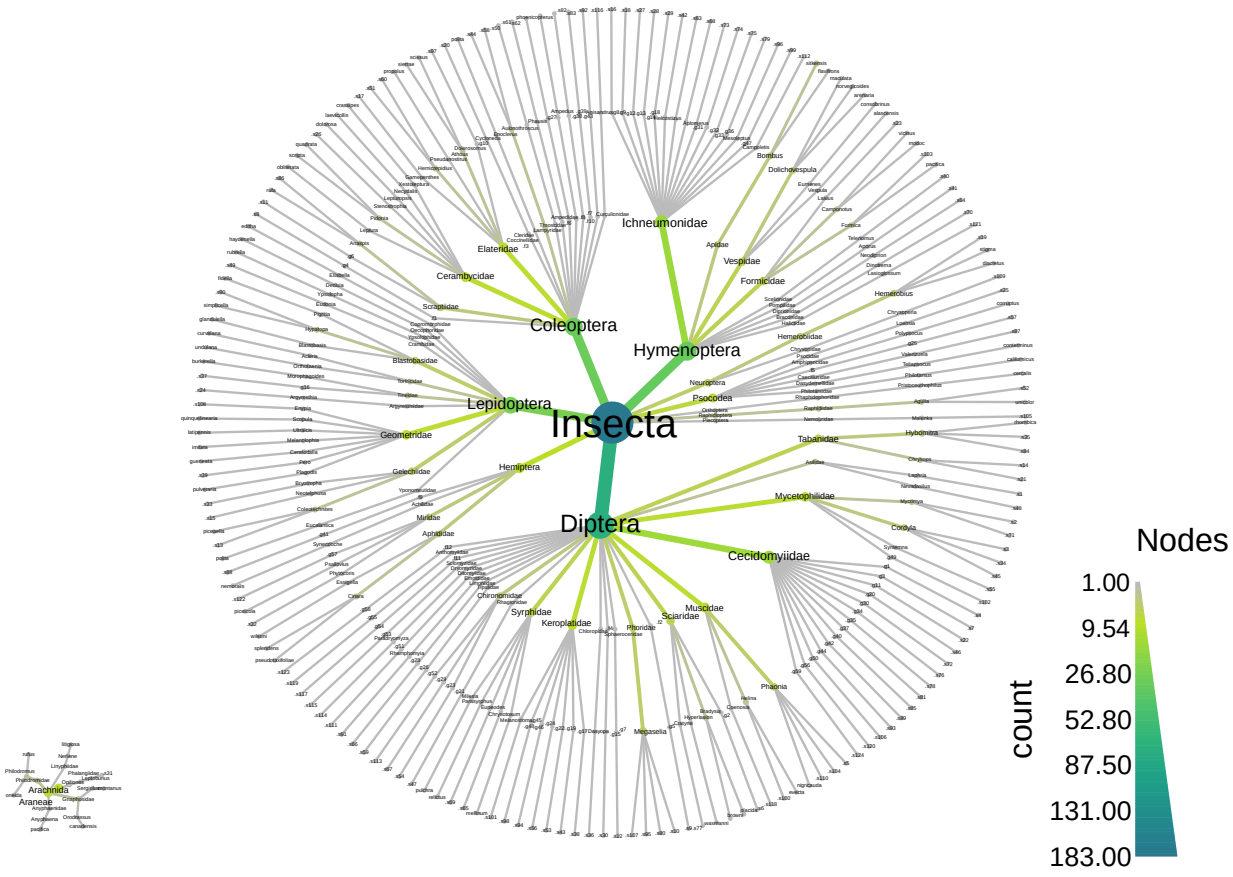


Figure 4S: Detailed taxonomic distribution of 190 Operational Taxonomic Units (OTUs) over two heat trees, the Insecta and the Arachnida. Node size and color are scaled to the number of OTUs in that node. Missing taxonomic information of species are indicated by the combination of a point, f, g or s, representing family, genus or species, respectively, and a number, e.g. '.f15'.

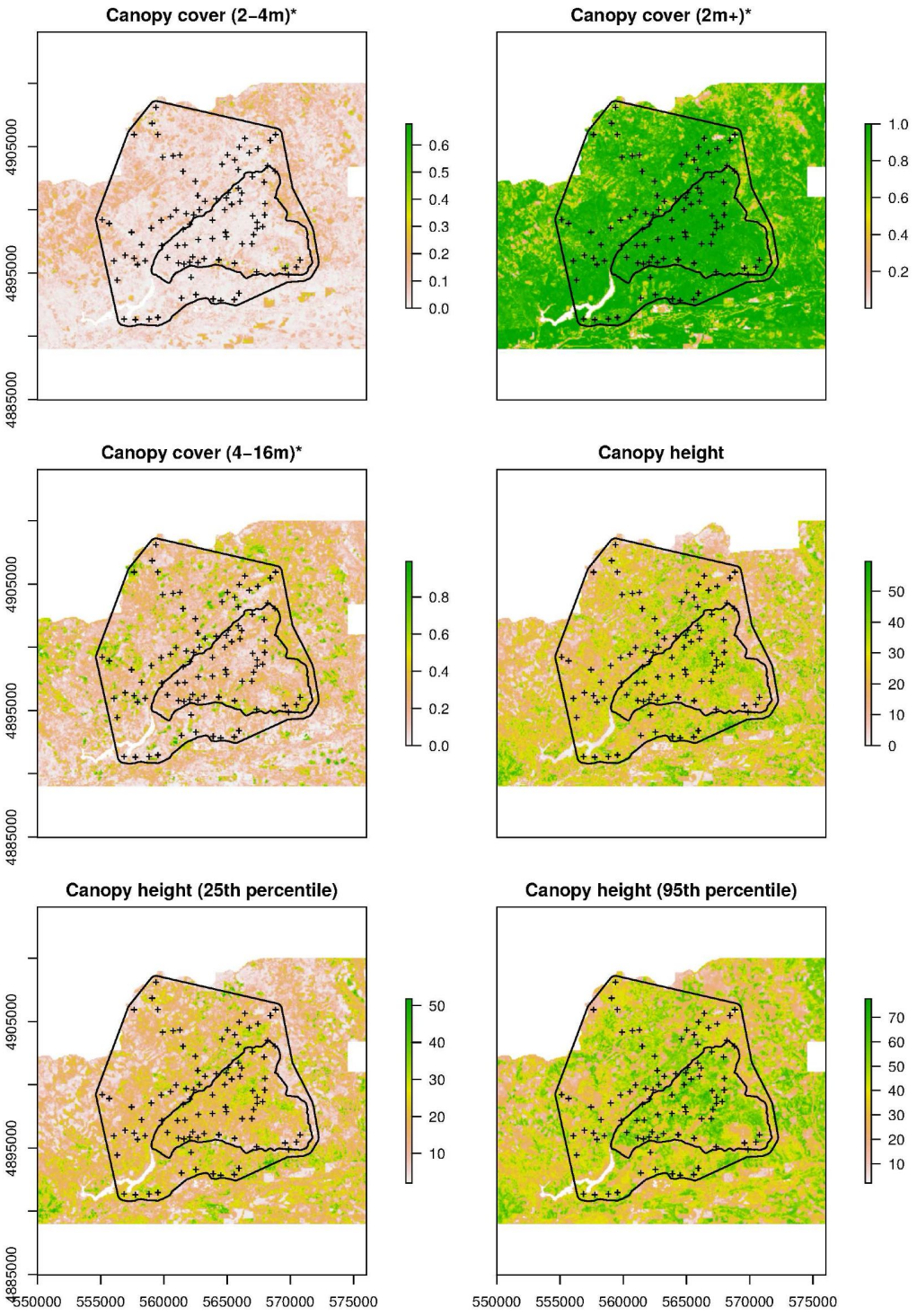


Figure 5S: All candidate covariates. Sample locations are marked by the plus sign, inner black outline shows H.J. Andrews Experimental Forest boundary and outer black outline shows extent of prediction area, Covariates used in model are marked with an asterisk. See Table S-covariates for covariate descriptions. The full figures are in `https://github.com/chnpenny/HJA_analyses_Kelpie_clean/blob/main/05_supplement/Plots/Figure_5S-full.pdf`.

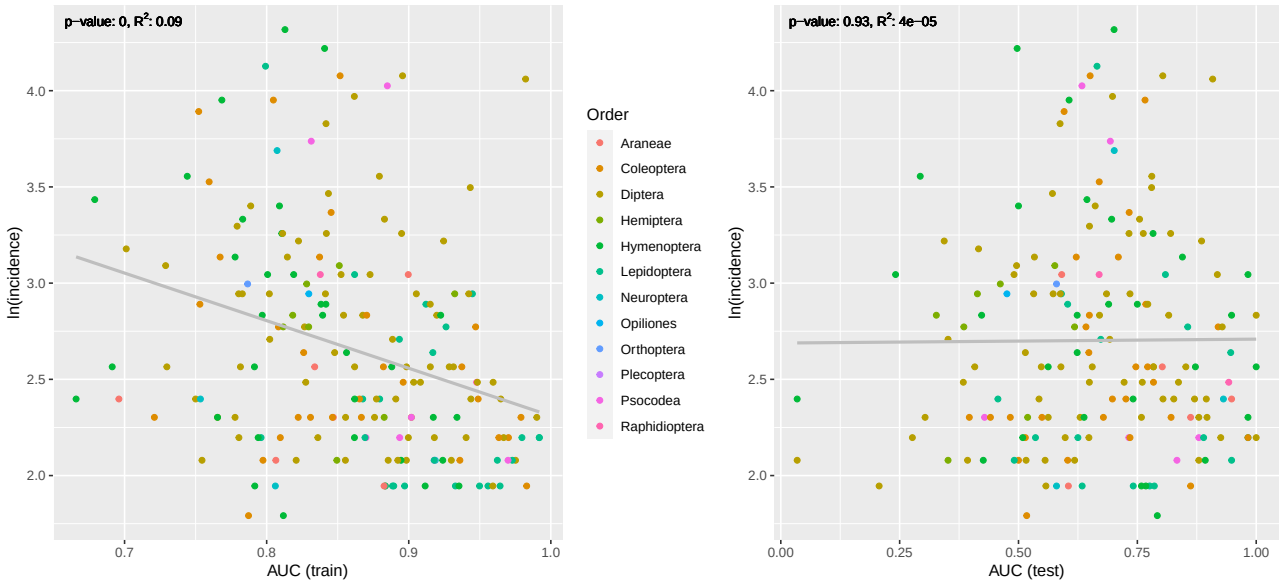


Figure 6S: Explanatory (training) and predictive (test) AUCs of all OTUs by incidence. Colors correspond with the order of OTUs. OTUs that are detected less (low incidence) show larger variance in the AUC values. The p-value and R^2 of the linear regressions are shown on the top of the plots. To be noticed, incidences of OTUs are log-transformed.

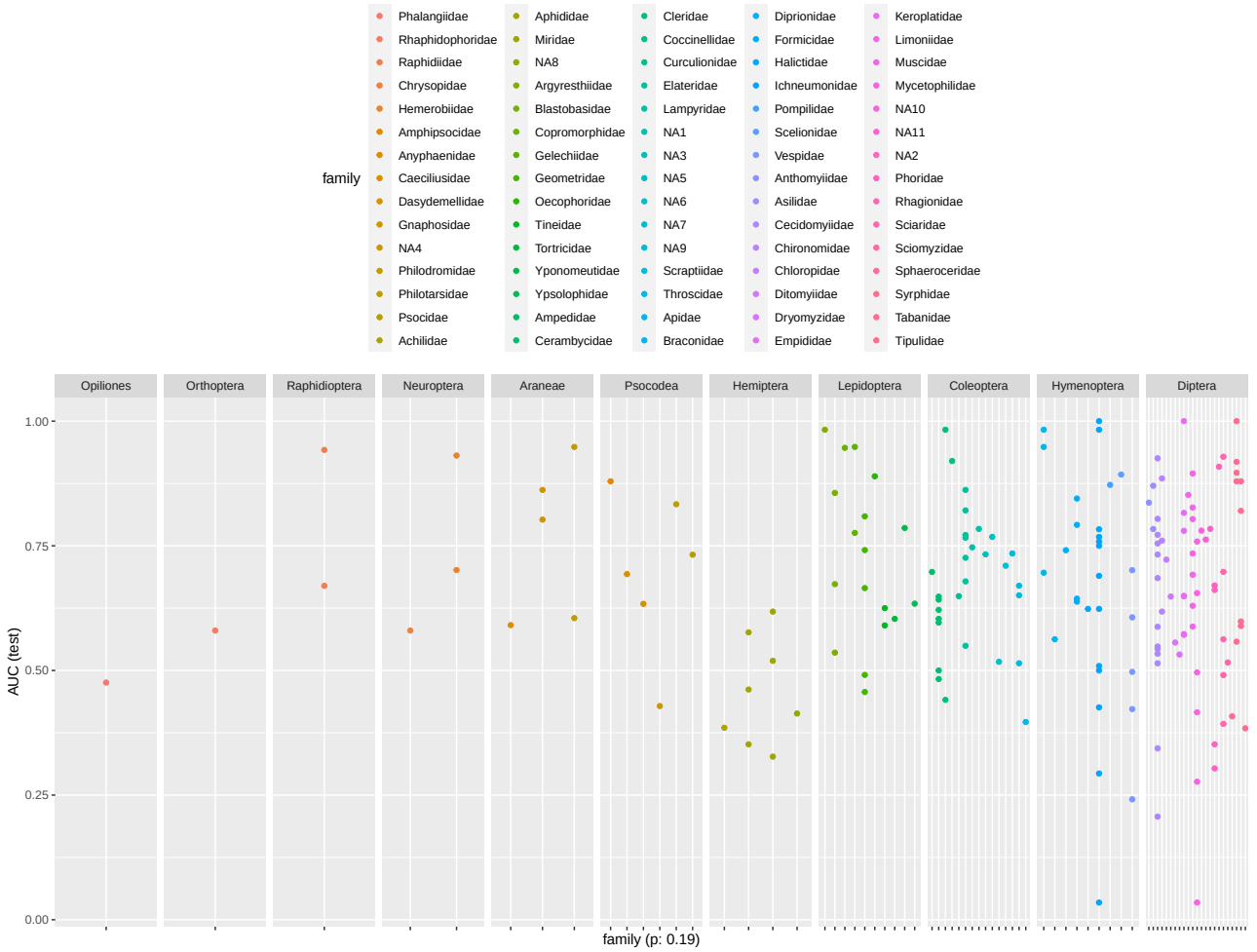


Figure 7S: Predictive AUCs of all OTUs by taxonomic family. Colors correspond with the family information and they are arranged according to the order information. A linear regression shows that there is no significant effect of family on the predictive AUCs (p-value 0.19 for this regression).

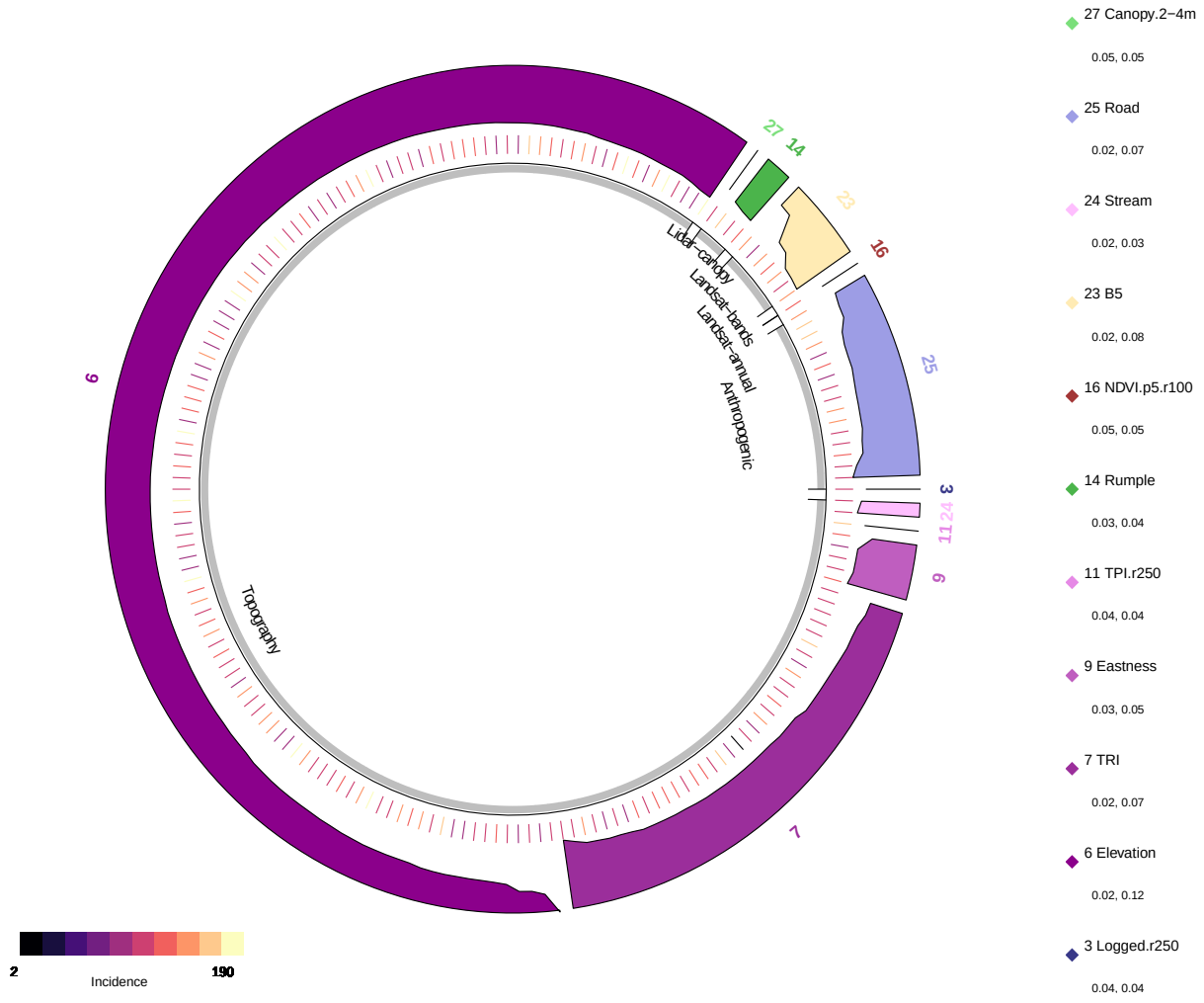


Figure 8S: The most important environmental covariate with regard to interaction effects for each OTU, excluding spatial location variables. Each tick mark on the middle ring represents an OTU, coloured by its incidence (see legend lower left), with the outer colour bands indicating its most important individual covariate from the point of view of interaction strength. The effect of environmental covariates on species (i.e. OTU) distributions is comprised of its individual effect and its effect through interacting with other covariates (detail in section, Variable importance with explainable AI (xAI)). Gray bands in the inner ring indicate covariate groupings (Table 1S). Elevation (variable 6) and TRI (variable 7) are the most important variables for the most OTUs. The heights of the colour bands are scaled to the Friedman's H statistic for overall interaction strength for that OTU. The ranges of overall interaction strength for each environmental variable are shown in the legend on the right.

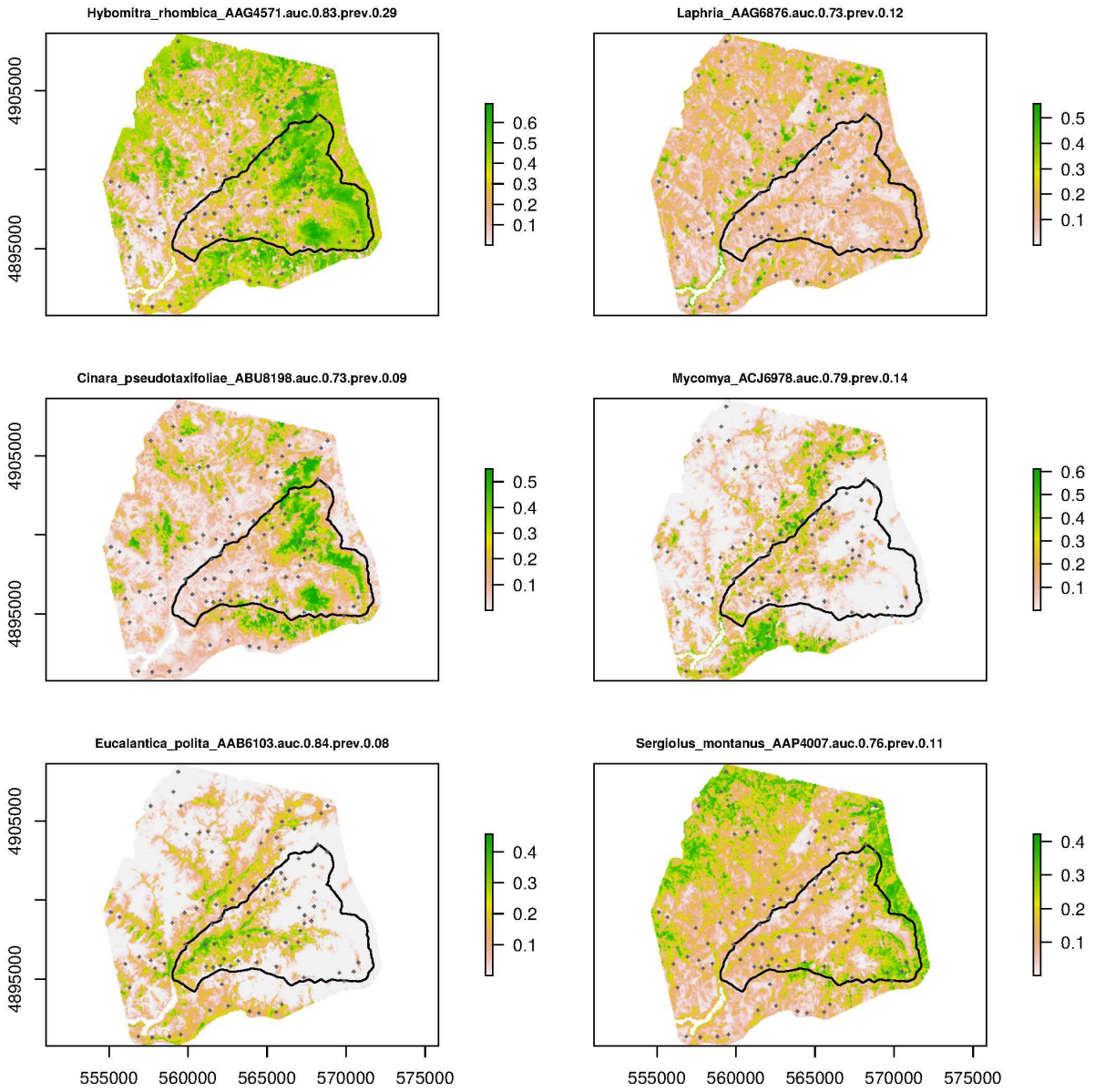


Figure 9S: Individual, interpolated species distributions. The full figure is in `https://github.com/chnpenny/HJA_analyses_Kelpie_clean/blob/main/05_supplement/Plots/Figure_9S-full.pdf`

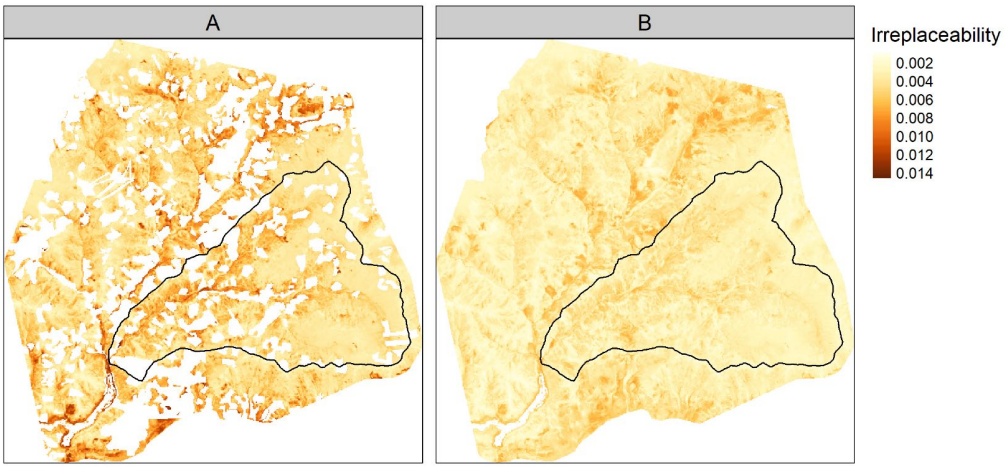


Figure 10S: Site-irreplaceability values plotted across the study area, showing HJA Experimental Forest boundaries (black line). A. With plantations masked out. B. With plantations present. Note the higher irreplaceability values in the unmasked part of the landscape (mostly along stream courses), which is because the species that are mainly restricted to plantations are rarer across our study area than those in old growth forests.

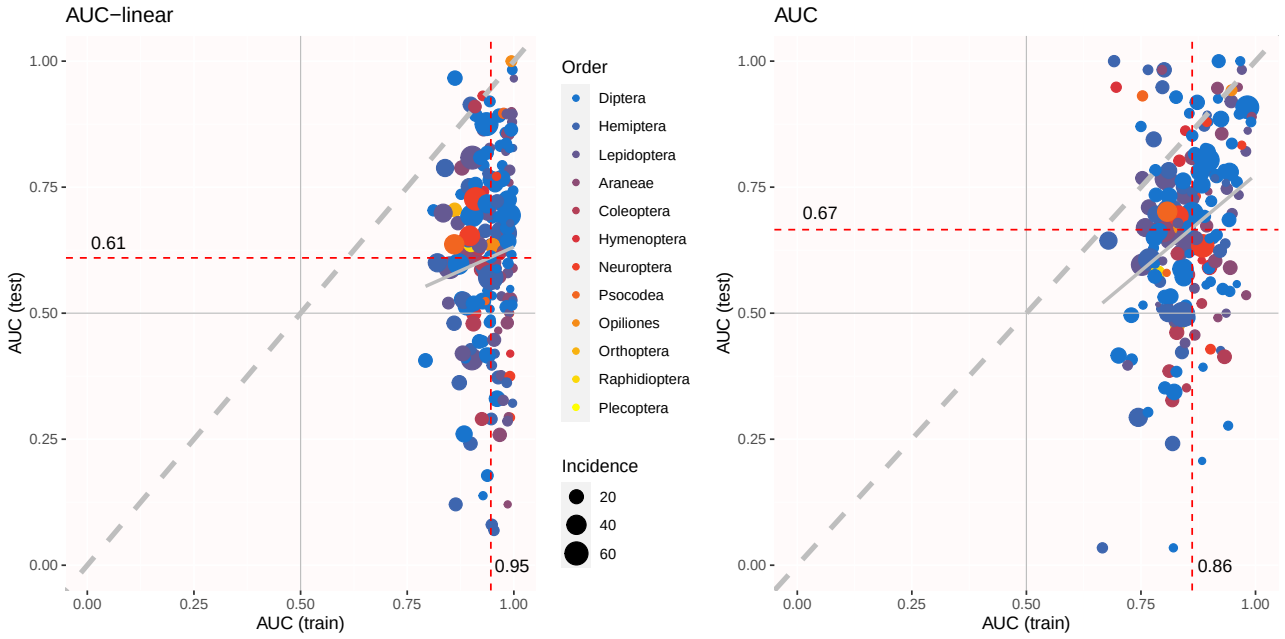


Figure 11S: Explanatory and predictive AUCs of the tuned sjSDM model applying linear fitting on the environmental part (left panel) to the same model applying DNN fitting (right panel). The explanatory power (x axis, AUC (train)) is higher but the predictive power (y axis, AUC (test)) is lower in the linear model, relative to the DNN model.

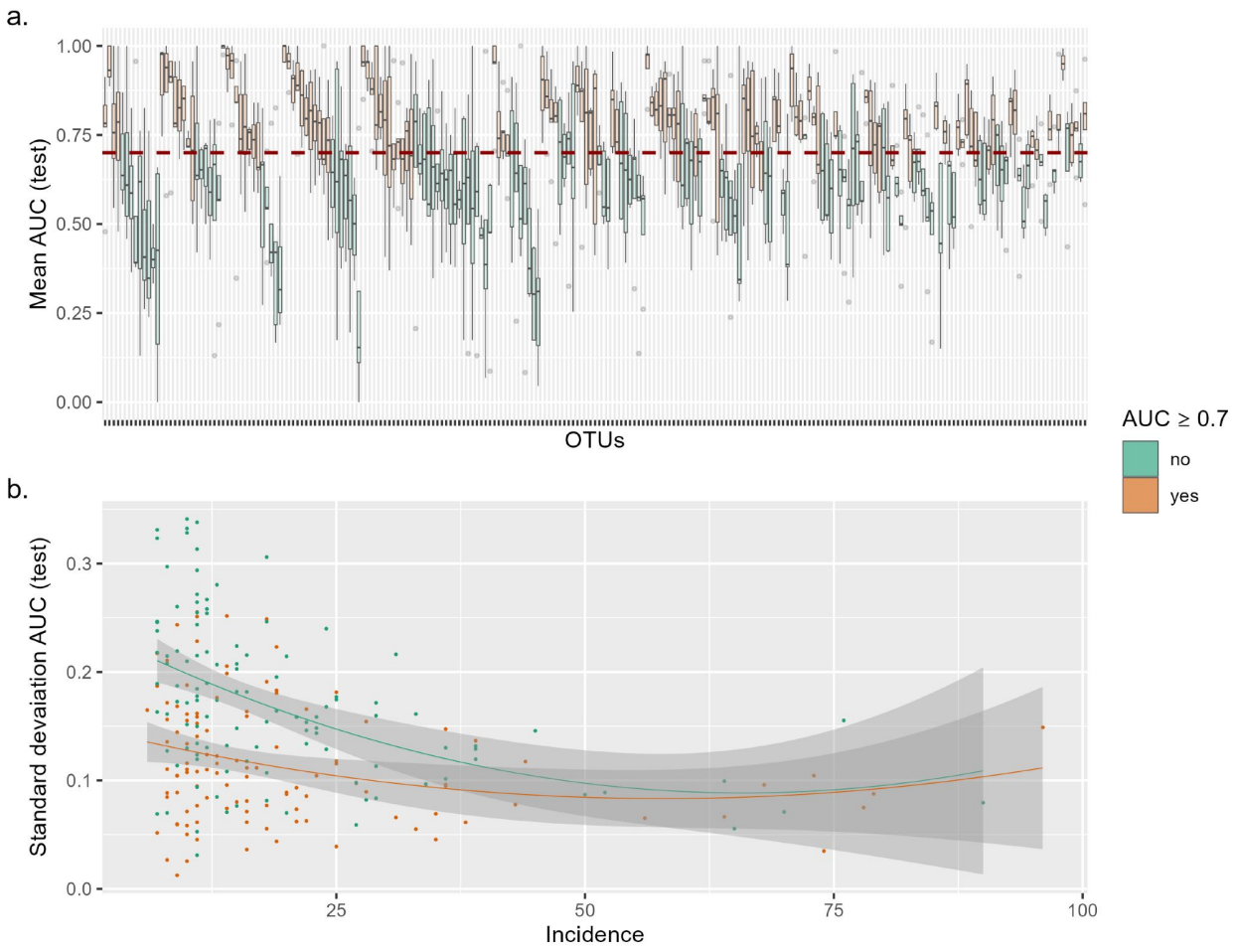


Figure 12S: Variability in AUC scores for all OTUs as evaluated with 5-fold cross validation. Variability in AUC is only weakly higher for lower incidence OTUs, and mean AUC does not increase with higher incidence. a) OTUs (boxes) in orange have $AUC_{mean} \geq 0.70$, and those in green have $AUC_{mean} < 0.70$. OTUs are ordered by increasing incidence, from occurrence at 6 sample points (far left) to occurrence at 96 sample points (far right). The dashed red line is at $AUC = 0.70$, which is the threshold value for including OTUs in further analysis. b) Standard deviation of AUC as a function of incidence. Regression lines shown from a polynomial linear model on OTUs with $AUC_{mean} \geq 0.70$ ($R^2 = 0.05$, $p = 0.029$, $df = 2, 109$) and with $AUC_{mean} < 0.70$ ($R^2 = 0.19$, $p < 0.001$, $df = 2, 110$).

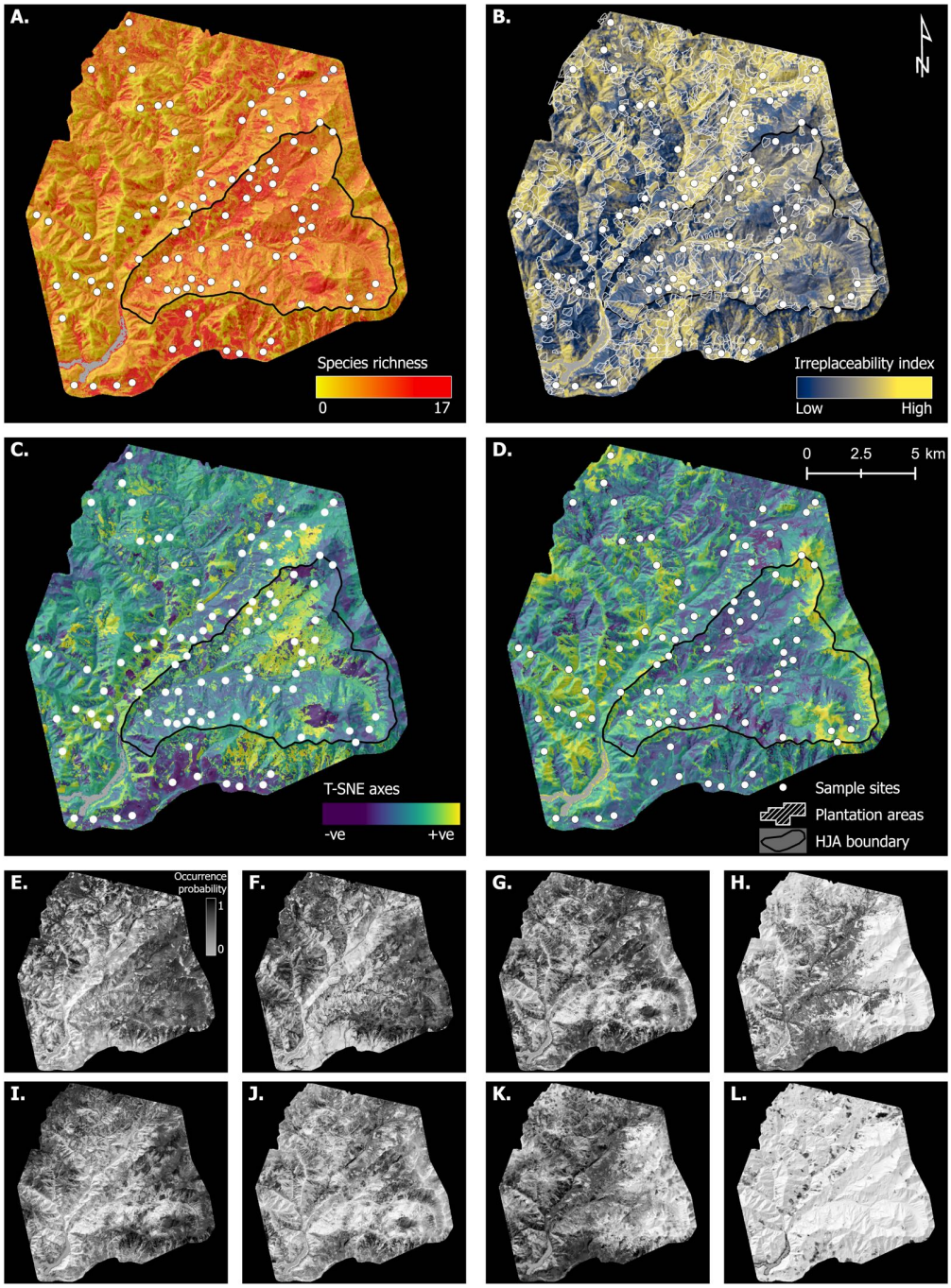


Figure 13S: Alternative version of Figure 2 (main text) using OTUs with $AUC_{mean} \geq 0.70$ from 5-fold CV on full data set (see methods in Tuning and Testing, above). JSDM-interpolated spatial variation in species richness, irreplaceability, and composition, plus examples of individual species distributions. 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-line-bordered triangular area delimits the H.J. Andrews Experimental Forest (HJA, see Figure 1, main text). E-L. Selected individual species distributions, with BOLD ID, predictive AUC, and prevalence. E. Rhagionidae gen. sp. (`BOLD: ACX1094`, AUC: 0.95, Prev: 0.64). F. *Plagodis pulveraria* (`BOLD: AAA6013`, AUC: 0.72, Prev: 0.23). G. *Phaonia* sp.(`BOLD: ACI3443`, AUC: 0.80, Prev: 0.65). H. *Orthotaenia undulana* (`BOLD: AAB4022`, AUC: 0.95, Prev: 0.06). I. *Helina evecta* (`BOLD: AAC2498`, AUC: 0.76, Prev: 0.16). J. *Diptera* sp. (`BOLD: AAZ4857`, AUC: 0.75, Prev: 0.16). K. *Blastobasis glandulella* (`BOLD: AAG8588`, AUC: 0.91, Prev: 0.18). L. *Dasyopa* sp. (`BOLD: ADI1308`, AUC: 0.82, Prev: 0.12)

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, Jana M. McPherson, Miguel B. Araújo, Roger Bivand, Janine Bolliger, Gudrun Carl, Richard G. Davies, Alexandre Hirzel, Walter Jetz, W. Daniel Kissling, Ingolf Kühn, Ralf Ohlemüller, Pedro R. Peres-Neto, Björn Reineking, Boris Schröder, Frank M. Schurr, and Robert Wilson. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628, 2007. doi: https://doi.org/10.1111/j.2007.0906-7590.05171.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2007.0906-7590.05171.x.

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.

William T. Langford, Ascelin Gordon, Lucy Bastin, Sarah A. Bekessy, Matt D. White, and Graeme Newell. Raising the bar for systematic conservation planning. *Trends in Ecology Evolution*, 26(12):634–640, December 2011. ISSN 0169-5347. doi: 10.1016/j.tree.2011.08.001.

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.

Szymon Maksymiuk, Alicja Gosiewska, and Przemyslaw Biecek. Landscape of R packages for eXplainable Artificial Intelligence. 2020. doi: 10.48550/ARXIV.2009.13248. URL https://arxiv.org/abs/2009.13248. Publisher: arXiv Version Number: 3.

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 (JSDM). *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.