Environmental DNA captures the internal structure of a pond metacommunity

Running title: Metabarcoding a Pond Metacommunity

Authors: Wang Cai[1,2], Maximilian Pichler[3], Jeremy Biggs[4,5], Pascale Nicolet[4], Naomi Ewald[4], Richard A. Griffiths[6], Alex Bush[7], Mathew A. Leibold[8], Florian Hartig[3], Douglas W. Yu[1,9,10,11*]

Author affiliations:

[1] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

[2] Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla 666303, China

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

[4] Freshwater Habitats Trust, First Floor, Bury Knowle House, North Place, Old High Street, Headington, Oxford, UK

[5] Oxford Brookes University, Headington, Oxford, OX3 0BP, UK

[6] Durrell Institute of Conservation and Ecology, School of Anthropology and Conservation, University of Kent, UK

[7] Lancaster Environment Centre, Lancaster University, UK

[8] Department of Integrative Biology, University of Texas at Austin, Austin, TX USA

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

[10] Yunnan Key Laboratory of Biodiversity and Ecological Security of Gaoligong Mountain, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

[11] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, UK

[*]**Correspondence:** email: dougwyu@mac.com; address: School of Biological Sciences, Univ. of East Anglia, Norwich, Norfolk NR4 7TJ UK

KEYWORDS: Aquatic eDNA, metabarcoding, novel community data, biodiversity, pond, joint species distribution model (JSDM), macroecology, *Triturus cristatus*

Abstract

Metacommunity theory unites community and spatial ecology. Recent innovations in sequencing technology may now allow us, for the first time, to confront this theory with entire trophic community data at large scales. We metabarcoded vertebrate eDNA from 320 ponds in England. Using this 'novel community data', we calculated the 'internal structure' of this metacommunity, which is the contributions of environmental, biotic, and spatial processes, as well as stochasticity, to the distribution of each species and the composition of each site. Our results suggest that community assembly processes in this pondscape do not conform to a single metacommunity paradigm, but rather vary among species and over space. Combining this variation with external knowledge of species biologies allows us to better understand the differential influences of environmental filtering, species interactions, and dispersal limitation on several of the vertebrate species, paving the way for the routine empirical study of metacommunities.

Introduction

Metacommunity theory, which explicitly models feedback between local communities and regional species pools, has been proposed as a unifying theory of spatial community ecology (Leibold *et al.* 2004; Leibold & Chase 2018). In a metacommunity, local communities are linked by dispersal. Community dynamics are further governed by environmental filtering, interactions among species, and ecological drift, which together determine the species composition of each local community (Vellend 2016).

Early metacommunity theory tended to study subsets of these four forces (dispersal, environmental filtering, species interactions, and drift), resulting in four paradigms of community assembly: species-sorting, mass-effect, patch-dynamic, and neutral communities (Holyoak *et al.* 2005; Shoemaker & Melbourne 2016; Suzuki & Economo 2021; Thompson *et al.* 2020), with most empirical studies focusing on the first two (Fávero *et al.* 2022; Logue *et al.* 2011; Urban 2004). The species-sorting paradigm (Van der Gucht *et al.* 2007) occurs, for example, if most species exhibit resource-use specialisation and can disperse freely over a heterogeneous environment. In this case, species sort by their niches, resulting in local community compositions determined by local environmental conditions. Mass effects (Rimet *et al.* 2023) are expected if some species have high dispersal ability and large population sizes, allowing them to immigrate into a local community and extirpate resident species, even if the resident species are better adapted to the local environment. The patch-dynamic paradigm (Yu *et al.* 2001) could arise if species exhibit the same environmental requirements but nonetheless coexist stably via different dispersal strategies over large-scale environmental heterogeneity, with or without a competitive hierarchy. Finally, the neutral and nearly-neutral paradigms (He 2005) arise in the absence or near-absence of species differences when species can coexist for a long time with abundances varying stochastically.

Undoubtedly, no real metacommunity conforms completely to any one paradigm, since all four forces are always acting. An ideal empirical analysis of metacommunity data would thus try to measure all four forces and their relative contributions. However, summary community metrics, such as alpha and beta diversities (Figure 1), have limited power to explain metacommunity structure (Guzman *et al.* 2022; Ovaskainen *et al.* 2019), in part because summary metrics cannot see the contributions of individual species and sites. This is why Leibold et al. (2022) refer to the summary-metric approach as studying the 'external structure' of metacommunities. External-structure analyses are probably more useful for inferring processes that affect large

numbers of species and sites similarly, but by their nature, they are not able to represent differences between species and locations in the relative strength of community assembly processes (Leibold *et al.* 2022).

Leibold *et al*. (2022) have thus proposed to expose the 'internal structure' of metacommunities, by using a joint species distribution model (JSDM) to partition the varying contributions of environmental filtering, species interactions, and dispersal limitation to the distributions of individual species and sites (Figure 1). One potential advantage is that views of the internal structure allow one to relate species differences to their trait differences and to relate site differences to their condition differences, thereby potentially linking pattern to process. Leibold *et al*. used simulation studies to show that internal-structure analysis can indeed differentiate synthetic metacommunities that differ in site environmental distinctiveness and in species niche breadth, dispersal ability, niche centrality, and the presence or absence of competitive interactions (Figure 1).

More elaborate study designs such as time series should further increase power in diagnosing the drivers of metacommunity structure, by providing ways to infer causality (Hartig *et al.* 2023). For instance, Terry *et al*. (2023) simulated metacommunities to show that a press disturbance (elevated temperature) reduces the variation attributable to biotic covariances in a JSDM, which they hypothesise is caused by species distributions changing at different rates, thereby continually creating new species co-distributions. As an empirical example, Pandit *et al*. (2009) used temporal data from natural microcosms to show that habitat specialists respond primarily to environmental factors, whereas habitat generalists respond primarily to spatial factors.

That said, synthetic metacommunities have the advantages of being both complete and simpler than real metacommunities. Leibold *et al*.'s synthetic sites varied along only one niche axis, their synthetic species only engaged in competition, and the species compositions in each site were completely measured. In contrast, real metacommunity datasets have detection failures, measurement errors, and model uncertainty; not all species, environmental covariates, and sites can be included; species interact in multiple ways; and real metacommunities might be non-stationary, not least because of climate change (Abrego *et al.* 2021; Terry *et al.* 2023).

Therefore, here we test whether it is possible to detect signals of the internal structure of a candidate real metacommunity, and we ask if the revealed structure matches expectations and what methodological improvements can be made for future studies. An empirical

metacommunity dataset ideally consists of inventories of many local communities with standardised species presence-absence or abundance information, (2) over an area connected (and large) enough for dispersal (and dispersal limitation) to operate, (3) with the taxonomic breadth to include interacting guilds such as predators and prey, and (4) measures of local environmental conditions relevant to the niche requirements of all these species.

To generate such a dataset, we use a set of ponds from England, UK, that were originally sampled for aquatic environmental DNA (eDNA) to detect the great crested newt (*Triturus cristatus*), a UK-protected amphibian species that breeds in ponds (Biggs *et al.* 2015). We metabarcoded the pondwater samples to detect vertebrate eDNA and generated a community matrix of 320 ponds × 48 vertebrate species. Each pond was associated with 8 environmental covariates and a geographical location, which let us fit three data matrices in a joint species distribution model (JSDM, Figure 1).

eDNA metabarcoding is one of a set of technological innovations that now allow ecologists to routinely generate 'novel community data' (reviewed in Hartig *et al.* 2023), but it is an open question how best to extract ecological information from this data. The internal structure approach seems a promising avenue, and the first such attempt with eDNA data, to our knowledge, is Vass *et al*. (2022), who metabarcoded aquatic fungi from four separated inlets along the Swedish coast and found that most species and site variance in a JSDM was attributable to the environmental covariates. This suggests that species sorting is the predominant assembly process and is consistent with Vass *et al*.'s *external*-structure analysis showing that sites cluster primarily by month. However, *within* month and inlet, the other assembly processes might act more strongly.

Here we report that our metabarcoded pondscape is consistent with Leibold *et al*.'s (2022) prediction that "species and sites can vary widely in how they contribute to…metacommunity-wide attributes" rather than conforming to any one metacommunity paradigm. Moreover, we show that the fitted environmental covariates are consistent with known fish and amphibian ecologies, that the negative biotic covariances between fish and amphibians are consistent with expected predator-prey relationship between the two taxa, and that high spatial autocorrelation for the palmate newt is consistent with a hypothesised relictual distribution for this species. On the other hand, environmentally distinctive sites are not more likely to be better explained by environmental covariates, although many positive biotic covariances suggest the action of environmental covariates that remain unmeasured.

Methods

Environmental covariates

To quantify the land cover around each pond, we used the Rowland et al.'s (2017) classification of satellite data into 21 UK land cover classes. For each pond, we calculated the proportions of land-class within a 500 m radius of its point location and used principal component analysis in the {FactoMineR} v.2.4 R package (Lê *et al.* 2008) to extract the top three principal components (accounting for 40% of the total variation, Figure S1, Table1). Each pond was also scored during sampling for ten standard pond variables used by surveyors to calculate the pond's Habitat Suitability Index (HSI) for the great crested newt (ARG-UK 2010), we used five of them (Table 1).

Pond sampling and newt assays

The pond water samples were the result of a single season's systematic great crested newt (GCN) survey of 544 ponds in the South Midlands of England, UK in 2017. The survey used a grid design to generate single-species presence-absence data as input to a species distribution model that underpins a biodiversity offset market (Box 3 in Hartig *et al.* 2023). Samples were collected and processed following Biggs et al. (2015), and were stored at ambient temperature until shipped to a commercial lab (NatureMetrics, Egham, UK), and DNA was extracted using a precipitation protocol (Tréguier *et al.* 2014), after which each sample's DNA was cleaned and subjected to 12 separate qPCR tests. After the qPCR assays, the residual eDNA was stored at -80°C. We loosely estimate that the cost of sample collection and great crested newt assays, including effort getting permission to sample private properties, exceeded US$120,000. Pond eDNA collected for GCN thus has several advantages for metacommunity studies in that we have a standardised, professionally collected, commercially processed, and spatially contiguous set of eDNA samples, with standardised environmental covariates.

Metabarcoding assays

In 2019, the residual eDNA samples were retrieved and subjected to metabarcoding at NatureMetrics (PCR) and at Kunming Institute of Zoology (library prep), using a two-step protocol targeting a 73-110 bp fragment of 12S ribosomal RNA (Riaz *et al.* 2011). In the first step, 12 replicates per sample were amplified and then pooled to maximise the targeted sequence (Lahoz-Monfort *et al.* 2016). In the second step, three independent PCR replicates were performed on the pooled samples using twin-tagging (Yang *et al.* 2021; Zepeda-Mendoza

et al. 2016). The three independent (separately tagged) PCR replicates per sample in the second step were pooled into three approximately equimolar libraries for bead purification, library preparation, and sequencing on an Illumina HiSeq platform (PE150) at Novogene Tianjin, China.

We processed the raw sequence data with the modified DAMe bioinformatics pipeline of Cai et al. (2021), where we filtered out PCR and sequencing errors by retaining only sequences present in ≥2 PCR replicates with a minimum copy number of 25 per PCR. After sequence clustering, we generated a table of 540 ponds by 74 OTUs (operational taxonomic units). We assigned taxonomies to the OTUs using PROTAX (Axtner et al. 2019; Somervuo et al. 2017), setting prior probabilities to 0.90 for a list of expected UK vertebrate species (Harper et al. 2018).

Joint species distribution modelling

To fit JSDMs to the observed community data, we converted OTU read counts to presence-absence data. We retained only OTUs present in ≥5 ponds and only sites with ≥1 targeted OTU (= vertebrate species present in the United Kingdom), which reduced the number of OTUs from 74 to 48 and the number of ponds from 540 to 320. We assigned species-level taxonomies to OTUs that received ≥98% PROTAX probability of species assignment, and we classified the OTUs into 6 trait groups: fish, amphibians, perching birds, waterfowl, mammals, and domestic species. Domestic species are whose distributions we deemed as determined largely by humans (Table 2).

We fit our data with two distinct JSDM structures. One model was fit to all the species in our dataset, terrestrial and aquatic (320 ponds × 48 species), and the other model fit only to the aquatic species (amphibians and fish, 279 ponds × 15 species; the fewer ponds are because we excluded those without aquatic species). *A priori*, aquatic species should be more likely to be filtered by pond characteristics, which make up 5 of our 8 environmental covariates. Thus, we expect the aquatic species, especially the fish, to act more like narrow-niche species (Figure 1C, top row), and the terrestrial species to act more like broad-niche species (Figure 1C, bottom row).

All models were fitted using the {sjSDM} package v.1.0.6 (Pichler & Hartig 2021) running under R 4.2.2 (R Core Team 2022). We used a binomial likelihood and a multivariate probit link, linear main effects for the eight environmental covariates, and a DNN (deep neural net)

spatial model. To avoid overfitting, a light elastic net regularisation (Zou & Hastie 2005) was applied to all regression slopes and weights of the DNN. The model fitting details are in Supplementary Materials.

Internal structure of the metacommunity

After fitting the two models (terrestrial + aquatic, aquatic), we used the ANOVA functions implemented in sjSDM (based on Leibold *et al.* 2022) to partition the variation of each species' distribution and each site's composition across environmental covariates (E), spatial autocorrelation (S), and biotic covariances (C). The results are visualised using ternary plots, where the positions of the species and sites reveal the relative contributions of E, S, and C, i.e. the metacommunity's internal structure viewed over species and sites. The explanation of E, S and C in a ternary plot is given in Figure 1B2.

Leibold et al. (2022) found that more environmentally extreme sites (at the ends of the single niche axis used) received higher E values. To test this result in a natural metacommunity, we scored the environmental distinctiveness of each pond from zero to one as the sum of the absolute values of the pond's eight covariates, after scaling each covariate from -1 to 1.

Model generality

Our results (Figures 2, 3, 4) are the outputs of a complex model that includes a linear environmental structure with eight environmental covariates and a deep neural network spatial structure with 30 x 2 layers. Complex models run a risk of overfitting, so to estimate the risk of overfitting after elastic-net regularisation, we performed a 20-fold cross-validation test using stratified multi-label sampling to partition the data (Gunopulos *et al.* 2011; Szymański & Kajdanowicz 2017). In each of 20 rounds, 19 folds (304 sites) were used as a training dataset, and the fitted model was used to predict species compositions in the hold-out fold (16 sites). Model fitting used the same structure, regularisation strengths, and scaled environmental and spatial covariates as the original full model. For each round and species, we calculated from the training dataset an explanatory-performance AUC (Area Under the Curve) metric. We also used each hold-out fold to calculate a predictive-performance AUC for each species: a measure of model generality. The final explanatory- and predictive-performance AUCs per species were the means over all 20 folds. Species with higher predictive AUCs are those whose fitted models are more general (Figures 2, 3, 4). See Supplementary Materials for more details.

Results

Internal structure of the pond metacommunity

When analysing aquatic species only, we find that the fish are bimodally arrayed along the E-C axis, with four species relatively better explained by environmental covariates (higher E values), and six species relatively better explained by biotic covariances (higher C values), which reflect the effects of unknown environmental covariates plus possible species interactions (Figure 2A upper left, 2B).

In contrast, none of the five amphibian species shows a high contribution of either environmental covariates or biotic covariances, but relative to fishes, amphibians show a marginally greater contribution of dispersal limitation (higher S values), especially the palmate newt (*Lissotriton helveticus*), whose distribution is mostly explained by dispersal limitation. This species was detected in 16 ponds, in three separate sections of the survey area (Figure S4), and its high S value indicates that other ponds in which this species was not detected have environmental conditions similar to those in which the palmate newt was detected.

Including terrestrial species in the analysis (Figure 2A lower left) has no qualitative effects on the results for fish species, whereas amphibian species show a higher relative contribution of biotic covariance, which could reflect either contributions of species interactions with terrestrial species or more unmeasured environmental covariates that are shared between amphibian and terrestrial species (Figure 2C). The terrestrial species themselves also largely range along the E-C axis, with no clear clustering by trait group. Like the aquatic species, only one terrestrial species, Mandarin duck (*Aix galericulata*), has a high S value (Figure S4) and is found in only five ponds.

For the two site ternary plots (Figure 2A, right column), the general effect of adding terrestrial species is an increase in the variance accounted for by biotic covariances (site points shift upward toward C). Accordingly, site R^2 is positively correlated with site species richness (linear model, $R^2 = 0.38$, $P < 0.001$, Figure S2). We tested Leibold et al.'s (2022) result that when species have narrow niche breadths, more environmentally extreme sites in their synthetic metacommunity have more of the variation in their compositions explained by environmental covariates (Figure 1C upper right). However, with our eight environmental covariates (Table 1), we found no evidence for this effect (general linear model with negative

binomial error; site-E-score ~ environmental distinctiveness; aquatic species: Tjur's $R^2 = 0.00$, $y = -0.8651 + -0.1812x$; all-species: Tjur's $R^2 = 0.00$, $y = -1.0979 + 0.1212x$; Figure S3).

We now examine species and site variation to try to infer some of the processes that have resulted in these observed internal structures.

Estimated environmental preferences

In the aquatic-only model (Figure 3A), the pond effects for the fish species are in the direction of greater prevalence in *larger* ponds with *lower* risk of drying and *less* macrophyte cover. Several of the fish species are known to eat macrophytes, reduce macrophyte cover through other behaviours, and/or require higher oxygen with less macrophyte cover (Lopes *et al.* 2015; Maceda-Veiga *et al.* 2017; Stefanoudis *et al.* 2017). In contrast, for the amphibians, pond effects are in the direction of greater prevalence in *smaller* ponds with *higher* macrophyte cover. Pond drying risk, water quality, and shade showed essentially no effects on amphibian prevalence. The effect of bordering land-cover varies across fish species, but for amphibians, the only effects are in the direction of greater prevalence of palmate newt and common frog in ponds bordered by woodland. All five amphibian species may utilise woodland outside of the breeding season (Mulkeen *et al.* 2017; Oldham *et al.* 2000).

In the aquatic+terrestrial model (Figure 3B), the effects of the environmental covariates on fish and amphibians remain largely the same as in the aquatic-only model. For the terrestrial species, most of the significant environmental-covariate effects are shade (% of pond perimeter shaded by trees), macrophyte cover, and land cover. Shade, which affects many perching birds and the grey squirrel, is most parsimoniously interpreted as increasing species detectabilities. Macrophyte cover has positive effects on three rodent species and the grey heron and negative effects on rabbits and ducks (Anatidae sp.). Land-cover effects are variable across species, but we note that cows and sheep have higher prevalences in ponds bordered by grassland, which is dominated by the 'improved grassland' subtype (Rowland *et al.* 2017), characterised by regular fertilisation, herbicide application and drainage (DEFRA & Rural Payments Agency 2022). Finally, we observe that the omnivorous European robin (*Erithacus rubecula*) and Eurasian coot (*Fulica atra*) both have higher prevalences at ponds of higher water quality; high water quality in the HSI scoring system is inferred from the presence of an "abundant and diverse invertebrate community" (ARG-UK 2010).

Biotic covariances

We visualise all residual biotic covariances in pairwise correlation plots (Figure 4), where large absolute correlation-coefficient values correlate with high C-values in the internal structure ternary plots (Figure 2A left column; linear model, aquatic species only, $R^2 = 0.726$, $P < 0.001$; aquatic + terrestrial, $R^2 = 0.201$, $P < 0.001$).

In the aquatic species model (Figure 4A, B), and after filtering to the 2.5% most negative and positive values, the only surviving negative correlations are between the common frog (*Rana temporaria*) and two omnivorous fish species (*Carassius carassius* and *Cyprinus carpio*) and between great crested newt (*Triturus cristatus*) and a carnivorous fish (*Esox lucius*). There are also three surviving positive correlations between fish species, which we conservatively interpret as indicating unmeasured environmental covariates.

In the aquatic + terrestrial species model (Figure 3C, D), and after filtering to the 2.5% most negative and positive values, most of the surviving positive correlations occur among the fish species and among the bird species, which we again interpret as unmeasured environmental covariates. There are now four surviving negative correlations between amphibians and fish. The common frog is negatively correlated with two omnivorous fish species (*Cyprinus carpio* and *Rutilus rutilus*) and great crested newt (*Triturus cristatus*) is negatively correlated with two carnivorous/omnivorous fish (*Esox lucius* and *Tinca tinca*). Also notable are negative correlations between several fish species with ring-necked pheasant (*Phasianus colchicus*) and with red fox (*Vulpes vulpes*).

Model generality

For the aquatic-only model, explanatory AUCs are always somewhat but not much greater than predictive AUCs, and explanatory and predictive AUCs are positively correlated (linear model, adjusted $R^2 = 0.546$, $P = 0.001$) (Figure 5A). For the aquatic+terrestrial model, explanatory AUCs are again still always greater than predictive AUCs, but the correlation weakens considerably (linear model, adjusted $R^2 = 0.139$, $P = 0.005$), and for some terrestrial species, the model makes worse-than-random predictions (predictive AUCs<0.5) (Figure 5B). The risk of overfitting is greater for low-predictive-AUC species, so the risk is therefore greater for the terrestrial species. Both models are generally better at predicting fish distributions than amphibian distributions. Palmate newt (*Lissotriton helveticus*) is an exception, with the overall second-highest predictive AUC (Figure 5A), which is probably explained by the large contribution of space to its explained variance. Stickleback (*Gasterosteus aculeatus*) and

European perch (*Perca fluviatilis*) are exceptions for fish, with the overall two lowest predictive AUCs in the aquatic model (Figure 5A).

Discussion

The goal of our study was to reveal the internal structure of a real metacommunity consisting of 320 systematically surveyed ponds with both resident and visiting species. We used a joint species distribution model to estimate the relative contributions of environmental covariates, spatial autocorrelations, and biotic covariances to explaining individual variation in species distributions and pond species compositions. Pondscapes are convenient study systems for metacommunity ecology because (1) each pond is unambiguously identified as a local community, (2) there is an *a priori* division between likely pond niche specialists (aquatic species) versus generalists (terrestrial species), and (3) aquatic eDNA metabarcoding can efficiently generate hundreds of community inventories, addressing the need for large and structured datasets, and the detected species encompass multiple trophic levels, increasing the possibility of detecting important species interactions.

Comparison of empirical metacommunity results with simulations

From the contributions of environmental covariates, biotic covariances, spatial autocorrelation, and unexplained variation (Figures 2, 3, 4, 5), we infer two internal structures, one for aquatic species only and one for aquatic + terrestrial species. That is, we infer the actions of environmental filtering, biotic interactions, dispersal limitation, and ecological drift on each species and site and, thus, how each species and site contributes to the apparent metacommunity. This gives us the first conclusion, which is that the internal structures are consistent with the four forces acting differently on different species and over space. A consequence is that no one metacommunity paradigm likely applies to this pondscape.

Inferring process from pattern

Our second conclusion is that these differences amongst species and sites appear to allow correct inference of the contributions of environmental filtering, dispersal limitation, and species interactions to the assembly of this candidate metacommunity. We base this conclusion on the plausibility of the patterns found when compared with our external knowledge of the species.

Environmental filtering. – Both of the inferred internal structures (Figure 2A left column) are more similar to the synthetic metacommunity with narrow niche breadths (Figure 1C upper row) than to the one with broad niche breadths (Figure 1C lower row). This suggests that environmental filtering is an important structuring force for some of the species in this pondscape. Accordingly, some of the environmental-covariate coefficient values are consistent with known biology. For instance, relative to amphibians, fish have a higher prevalence in larger, more permanent ponds (Figure 3). Macrophyte cover is also lower in ponds with higher prevalence of fish over amphibians, although the direction of causality might vary by species, since some fish species can directly reduce macrophyte cover. On the other hand, high macrophyte cover also appears to increase the prevalence of multiple terrestrial and waterfowl species (Figure 3). Finally, and contrary to its inclusion as a predictor of newt presence (ARG-UK 2010), higher water quality is not associated with higher prevalence of any amphibian species.

Dispersal limitation. – The contribution of space is dominant for only two species, palmate newt and mandarin duck (Figure 2). The spatial distribution of the 16 palmate newt detections is visibly patchy (Figure S4), and this species is observed to have a patchy distribution in the UK, is common in woodland, heathlands and meadows; in our case, the patchy distribution may be due to the persistence of populations with a historic distribution associated with woodland (Beebee & Griffiths 2000). There are only five detections for mandarin duck (Figure S4), so it is unclear whether such a spatial pattern implies dispersal limitation, but this introduced and free-living species is described as largely sedentary in the UK (animal diversity org. 2023).

Species associations (co-distribution). – With the main exception of common toad (*Bufo bufo*), which is protected from fish predation by toxins, residual associations between amphibians and fish are mostly negative, and three or four biotic correlations survive after filtering to the most extreme values (Figure 4). Keeping in mind the important caveat that biotic covariances alone should not be taken as direct evidence for species interactions (Blanchet *et al.* 2020; Dormann *et al.* 2018; Hartig *et al.* 2023; Poggiato *et al.* 2021; Zurell *et al.* 2018), it is still plausible that these associations arise from biotic interactions since the fish species involved are omnivorous or carnivorous and could thus prey on amphibian eggs, larvae, or possibly adults.

Site distinctiveness. – We tested Leibold et al.'s (2022) hypothesis that more environmentally distinctive sites (ponds) should be more strongly determined by environmental filtering but found no relationship between pond E-value and our measure of distinctiveness. However,

Leibold et al.'s simulations used only one niche axis, letting distinctiveness be unambiguously defined. In a real metacommunity, there are many contributing environmental covariates. As a result, (1) it is unclear how best to combine them into a single distinctiveness metric, (2) at least some of the species in our dataset are probably 'broad niche' with regards to pond characteristics, meaning that even environmentally distinct ponds are unlikely to result in distinct species compositions (Figure 1C lower row), and (3) some environmental covariates are unmeasured (Figure 4) and thus result in higher C-values for sites that are strongly filtered by those covariates. In this light, the increase in both site R^2 and C-value with increasing species richness (Figure S2) is consistent with the additional species adding ecological information, either in the form of unmeasured environmental covariates and/or true species-interaction information.

Combining metacommunity theory with novel community data

Taken together, our results support the hypothesis that internal-structure analysis can reveal plausible metacommunity processes from an observational dataset and test metacommunity hypotheses. Moreover, because novel community data (Hartig *et al.* 2023) enjoy large gains in efficiency and error homogeneity over traditional survey methods, which are prone to observer effects, the way is finally paved for the routine study of metacommunities. It is now also realistic to implement several study-design changes that can further improve the inference of metacommunity processes and to efficiently generate information that is useful for biodiversity management.

Sample replicates and detection models. – Sample eDNA was left over from great crested newt qPCR testing, eDNA was originally collected using ethanol precipitation instead of filters, which capture more eDNA (Bruce *et al.* 2021; Goldberg *et al.* 2016), and only one sample was taken per pond. These features all increase false-negative error rate. False-positive errors are also a risk given that a study of eDNA qPCR testing for great crested newts estimated a 5.2% false-negative and a 1.5% false-positive rate in the field (Buxton *et al.* 2021). Future metacommunity studies could be improved by combining a JSDM with sample replicates and a detection model to account for observation error (Diana *et al.* 2022; Guillera-Arroita *et al.* 2017; Hartig *et al.* 2023; Tobler *et al.* 2019). For instance, the positive effect of shading on passerine birds and the grey squirrel is most likely a detection effect (Figure 3).

Survey design. – Near-neighbour ponds were not sampled in this survey, making it impossible to detect fine-scale spatial autocorrelation, which probably reduced the perceived relative

importance of dispersal limitation. Survey designs should incorporate a spatial distribution of ponds that allows estimation of the contribution of spatial autocorrelation and biotic covariances at different scales. It is known that the strength and direction of biotic covariances in fitted JSDMs change with spatial scale, possibly reflecting the actions of different unmeasured environmental covariates acting at different scales (Dormann *et al.* 2018; Zurell *et al.* 2018). Other improvements include better estimates of travel-cost distance between sampled sites, a larger set of environmental covariates, time series data, which could improve inference on metacommunity processes (Abrego *et al.* 2021; Guzman *et al.* 2022; Hartig *et al.* 2023; Jabot *et al.* 2020), and inference of abundance-change information from eDNA data (Diana *et al.* 2022; Ershova *et al.* 2021; Hoshino & Inagaki 2017; Kelly *et al.* 2019; Luo *et al.* 2023).

Generating information for more effective biodiversity management. – Finally, novel community data and metacommunity theory can be powerfully combined to generate the high-quality, granular, timely, and trustworthy information that can make biodiversity management more effective (Dietz *et al.* 2003). For example, thousands of ponds are already sampled each year in England, funded by a great crested newt compensation scheme known as 'District Licensing' (Natural England 2019), and because water samples can be reliably collected through a citizen-science network (Biggs *et al.* 2015), it is possible to use eDNA metabarcoding and internal-structure analysis to generate large-scale time series on the status and niche requirements of many vertebrate and invertebrate species. Granular data on species distributions also serve as input for systematic conservation planning. A recent attempt to carry out a systematic conservation plan in southern England was forced to use habitat classes rather than species data because most English species records are restricted to urban and protected areas (Smith *et al.* 2021).

Acknowledgments

We are grateful to the Freshwater Habitats Trust for collecting the samples and to NatureMetrics for donating the extracted DNA. D.W.Y. was supported by the Chinese Academy of Sciences (CAS) Strategic Priority Research Program (XDA20050202) and the CAS Key Research Program of Frontier Sciences (QYZDY-SSW-SMC024). W.C. is supported by the Postdoctoral Fellowship of Xishuangbanna Tropical Botanical Garden, CAS, and the China Postdoctoral Science Foundation (2023M743747). M.A.L. is supported by NSF

award 2025118. This paper is an output of the sCom working group supported by sDiv, the Synthesis Centre of iDiv (DFG FZT 118, 202548816).

D.W.Y. is a cofounder of NatureMetrics, which provides commercial eDNA services. NatureMetrics is a minor shareholder in the NatureSpace Partnership. No further conflicts of interest to declare.

Statement of authorship

WC and DWY designed the study; WC performed the molecular experiments; WC and DWY performed the bioinformatic analyses; WC and MP performed the statistical analyses; MP and FH developed new sjSDM functions; AB provided code to estimate land cover; WC and DWY wrote the first draft of the manuscript; all authors contributed substantially to revisions.

References

Abrego, N., Roslin, T., Huotari, T., Ji, Y., Schmidt, N.M., Wang, J., *et al.* (2021). Accounting for species interactions is necessary for predicting how arctic arthropod communities respond to climate change. *Ecography*, 44, 885–896.

Animal Diversity org. (2023). *Aix galericulata* Mandarin duck. Available at https://animaldiversity.org/accounts/Aix_galericulata/, Last accessed 11 Dec 2023.

ARG-UK. (2010). *ARG UK Advice Note 5: Great Crested Newt Habitat Suitability Index*. Amphibian and Reptile Groups of the United Kingdom.

Axtner, J., Crampton-Platt, A., Hörig, L.A., Mohamed, A., Xu, C.C.Y., Yu, D.W., *et al.* (2019). An efficient and robust laboratory workflow and tetrapod database for larger scale environmental DNA studies. *GigaScience*, 8, giz029.

Beebee, T.J.C. & Griffiths, R.A. (2000). *Amphibians and reptiles: a natural history of the British Herpetofauna*. HarperCollins, London, The new naturalist. 1. publ.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R.A., *et al.* (2015). Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (*Triturus cristatus*). *Biological Conservation*, 183, 19–28.

Blanchet, F.G., Cazelles, K. & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecol Lett*, 23, 1050–1063.

Bruce, K., Blackman, R., Bourlat, S.J., Hellström, A.M., Bakker, J., Bista, I., *et al.* (2021). *A practical guide to DNA-based methods for biodiversity assessment*. Pensoft Publishers.

Buxton, A., Matechou, E., Griffin, J., Diana, A. & Griffiths, R.A. (2021). Optimising sampling and analysis protocols in environmental DNA studies. *Sci Rep*, 11, 11637.

Cai, W., Yang, C., Wang, X., Wu, C., Larrieu, L., Lopez-Vaamonde, C., *et al.* (2021). The ecological impact of pest-induced tree dieback on insect biodiversity in Yunnan pine plantations, China. *Forest Ecol and Manag*, 491, 119173.

DEFRA & Rural Payments Agency. (2022). *Guidance. The SFI improved grassland soils standard*. Available at: https://www.gov.uk/guidance/the-sfi-improved-grassland-soils-standard. Last accessed 19 March 2023.

Diana, A., Matechou, E., Griffin, J., Yu, D., Luo, M., Tosa, M., *et al.* (2022). eDNAPlus: A unifying modelling framework for DNA-based biodiversity monitoring. *arxiv.org*, https://arxiv.org/abs/2211.12213.

Dietz, T., Ostrom, E. & Stern, P.C. (2003). The Struggle to Govern the Commons. *Science*,

302, 1907–1912.

Dormann, C.F., Bobrowski, M., Dehling, D.M., Harris, D.J., Hartig, F., Lischke, H., *et al.* (2018). Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. *Global Ecol Biogeogr*, 27, 1004–1016.

Ershova, E.A., Wangensteen, O.S., Descoteaux, R., Barth-Jensen, C. & Præbel, K. (2021). Metabarcoding as a quantitative tool for estimating biodiversity and relative biomass of marine zooplankton. *ICES Journal of Marine Science*, 78, 3342–3355.

Fávero, A.A., Büneker, H.M. & Waechter, J.L. (2022). Patterns and metacommunity processes of epiphytic bromeliad assemblages along a coastal-inland gradient in a subtropical Brazilian geographic corridor. *Austral Ecology*, 47, 1184–1199.

Goldberg, C.S., Turner, C.R., Deiner, K., Klymus, K.E., Thomsen, P.F., Murphy, M.A., *et al.* (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods Ecol Evol*, 7, 1299–1307.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Rooyen, A.R., Weeks, A.R. & Tingley, R. (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods Ecol Evol*, 8, 1081–1091.

Gunopulos, D., Hofmann, T., Malerba, D. & Vazirgiannis, M. (Eds.). (2011). *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg.

Guzman, L.M., Thompson, P.L., Viana, D.S., Vanschoenwinkel, B., Horváth, Z., Ptacnik, R., *et al.* (2022). Accounting for temporal change in multiple biodiversity patterns improves the inference of metacommunity processes. *Ecology*, 103, e3683.

Harper, L.R., Lawson Handley, L., Hahn, C., Boonham, N., Rees, H.C., Gough, K.C., *et al.* (2018). Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (*Triturus cristatus*). *Ecol Evol*, 8, 6330–6341.

Hartig, F., Abrego, N., Bush, A., Chase, J.M., Guillera-Arroita, G., Leibold, M.A., *et al.* (2023). Novel community data in ecology-properties and prospects. *Trends Ecol Evol*, in press (accepted).

He, F. (2005). Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. *Funct Ecology*, 19, 187–193.

Holyoak, M., Leibold, M.A. & Holt, R.D. (2005). *Metacommunities Spatial Dynamics and Ecological Communities*. Univ. of Chicago Press, chicago.

Hoshino, T. & Inagaki, F. (2017). Application of Stochastic Labeling with Random-Sequence Barcodes for Simultaneous Quantification and Sequencing of Environmental 16S rRNA Genes. *PLoS ONE*, 12, e0169431.

Jabot, F., Laroche, F., Massol, F., Arthaud, F., Crabot, J., Dubart, M., *et al.* (2020). Assessing metacommunity processes through signatures in spatiotemporal turnover of community composition. *Ecol Lett*, 23, 1330–1339.

Kelly, R.P., Shelton, A.O. & Gallego, R. (2019). Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies. *Sci Rep*, 9, 12133.

Lahoz-Monfort, J.J., Guillera-Arroita, G. & Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Mol Ecol Resour*, 16, 673–685.

Lê, S., Josse, J. & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Soft.*, 25.

Leibold, M.A. & Chase, J.M. (2018). *Metacommunity ecology*. Princeton Univ. Press, Princeton, NJ.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., *et al.*

(2004). The metacommunity concept: a framework for multi-scale community ecology: The metacommunity concept. *Ecol Lett*, 7, 601–613.

Leibold, M.A., Rudolph, F.J., Blanchet, F.G., De Meester, L., Gravel, D., Hartig, F., *et al.* (2022). The internal structure of metacommunities. *Oikos*, 2022, e08618.

Logue, J.B., Mouquet, N., Peter, H. & Hillebrand, H. (2011). Empirical approaches to metacommunities: a review and comparison with theory. *Trends Ecol Evol*, 26, 482–491.

Lopes, T.M., Cunha, E.R., Silva, J.C.B., Behrend, R.D.L. & Gomes, L.C. (2015). Dense macrophytes influence the horizontal distribution of fish in floodplain lakes. *Environ Biol Fish*, 98, 1741–1755.

Luo, M., Ji, Y., Warton, D. & Yu, D.W. (2023). Extracting abundance information from DNA -based data. *Mol Ecol Resour*, 23, 174–189.

Maceda-Veiga, A., López, R. & Green, A.J. (2017). Dramatic impact of alien carp Cyprinus carpio on globally threatened diving ducks and other waterbirds in Mediterranean shallow lakes. *Biol Conserv*, 212, 74–85.

Mulkeen, C.J., Gibson-Brabazon, S., Carlin, C., Williams, C.D., Healy, M.G., Mackey, P., *et al.* (2017). Habitat suitability assessment of constructed wetlands for the smooth newt (*Lissotriton vulgaris* [Linnaeus, 1758]): A comparison with natural wetlands. *Ecol Eng*, 106, 532–540.

Natural England. (2019). *A Framework For District Licensing Of Development Affecting Great Crested Newts* (Natural England Technical Information Note TIN176 No. TIN176). Natural England.

Oldham, R.S., Keeble, J., Swan, M.J.S. & Jeffcote, M. (2000). Evaluating the suitability of habitat for the Great Crested Newt (*Triturus cristatus*). *Herpetol J*, 10, 143–155.

Ovaskainen, O., Rybicki, J. & Abrego, N. (2019). What can observational data reveal about metacommunity processes? *Ecography*, 42, 1877–1886.

Pandit, S.N., Kolasa, J. & Cottenie, K. (2009). Contrasts between habitat generalists and specialists: an empirical extension to the basic metacommunity framework. *Ecology*, 90, 2253–2262.

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.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J.S. & Thuiller, W. (2021). On the Interpretations of Joint Modeling in Community Ecology. *Trends Ecol Evol*, 36, 391–401.

R Core Team. (2022). R: A language and environment for statistical computing.

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P. & Coissac, E. (2011). ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Res*, 39, e145–e145.

Rimet, F., Canino, A., Chonova, T., Guéguen, J. & Bouchez, A. (2023). Environmental filtering and mass effect are two important processes driving lake benthic diatoms: Results of a DNA metabarcoding study in a large lake. *Mol Ecol*, 32, 124–137.

Rowland, C.S., Morton, R.D., Carrasco, L., McShane, G., O'Neil, A.W. & Wood, C.M. (2017). Land Cover Map 2015 (vector, GB). *NERC Environmental Information Data Centre. (Dataset)*.

Shoemaker, L.G. & Melbourne, B.A. (2016). Linking metacommunity paradigms to spatial coexistence mechanisms. *Ecology*, 97, 2436–2446.

Smith, R.J., Cartwright, S.J., Fairbairn, A.C., Lewis, D.C., Gibbon, G.E.M., Stewart, C.L., *et al.* (2021). Developing a nature recovery network using systematic conservation planning. *Conservat Sci and Prac*, 4, e578.

Somervuo, P., Yu, D.W., Xu, C.C.Y., Ji, Y., Hultman, J., Wirta, H., *et al.* (2017). Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods Ecol Evol*, 8, 398–407.

Stefanoudis, P.V., Sayer, C.D., Greaves, H.M., Davidson, T.A., Robson, H., Almeida, D., *et al.* (2017). Consequences of fish for cladoceran, water beetle and macrophyte communities in a farmland pond landscape: implications for conservation. *fal*, 190, 141–156.

Suzuki, Y. & Economo, E.P. (2021). From species sorting to mass effects: spatial network structure mediates the shift between metacommunity archetypes. *Ecography*, 44, 715–726.

Szymański, P. & Kajdanowicz, T. (2017). *A Network Perspective on Stratification of Multi-Label Data*. Presented at the Proceedings of Machine Learning Research, pp. 22–35.

Terry, J.C.D., Langdon, W. & Rossberg, A.G. (2023). Codistribution as an indicator of whole metacommunity response to environmental change. *Ecography*, 2023, e06605.

Thompson, P.L., Guzman, L.M., De Meester, L., Horváth, Z., Ptacnik, R., Vanschoenwinkel, B., *et al.* (2020). A process-based metacommunity framework linking local and regional scale community ecology. *Ecol Lett*, 23, 1314–1329.

Tobler, M.W., Kéry, M., Hui, F.K.C., Guillera-Arroita, G., Knaus, P. & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, 100(8), e02754.

Tréguier, A., Paillisson, J.-M., Dejean, T., Valentini, A., Schlaepfer, M.A. & Roussel, J.-M. (2014). Environmental DNA surveillance for invertebrate species: advantages and technical limitations to detect invasive crayfish *Procambarus clarkii* in freshwater ponds. *J Appl Ecol*, 51, 871–879.

Urban, M.C. (2004). Disturbance heterogeneity determines freshwater metacommunity structure. *Ecology*, 85, 2971–2978.

Van der Gucht, K., Cottenie, K., Muylaert, K., Vloemans, N., Cousin, S., Declerck, S., *et al.* (2007). The power of species sorting: Local factors drive bacterial community composition over a wide range of spatial scales. *Proc. Natl. Acad. Sci. U.S.A.*, 104, 20404–20409.

Vass, M., Eriksson, K., Carlsson-Graner, U., Wikner, J. & Andersson, A. (2022). Co-occurrences enhance our understanding of aquatic fungal metacommunity assembly and reveal potential host–parasite interactions. *FEMS Microbiology Ecol*, 98, fiac120.

Vellend, M. (2016). *The theory of ecological communities*. Monographs in population biology. Princeton University Press, Princeton.

Yang, C., Bohmann, K., Wang, X., Cai, W., Wales, N., Ding, Z., *et al.* (2021). Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *Methods Ecol Evol*, 12, 1252–1264.

Yu, D.W., Wilson, H.B. & Pierce, N.E. (2001). An empirical model of species coexistence in a spatially structured environment. *Ecology*, 82, 11.

Zepeda-Mendoza, M.L., Bohmann, K., Carmona Baez, A. & Gilbert, M.T.P. (2016). DAMe: a toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses. *BMC Res Notes*, 9, 255.

Zou, H. & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *J R Stat Soc Series B Stat Methodol*, 67, 301–320.

Zurell, D., Pollock, L.J. & Thuiller, W. (2018). Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography*, 41, 1812–1819.

Figures and legends

Figure 1. External-structure versus internal-structure analysis of metacommunities.

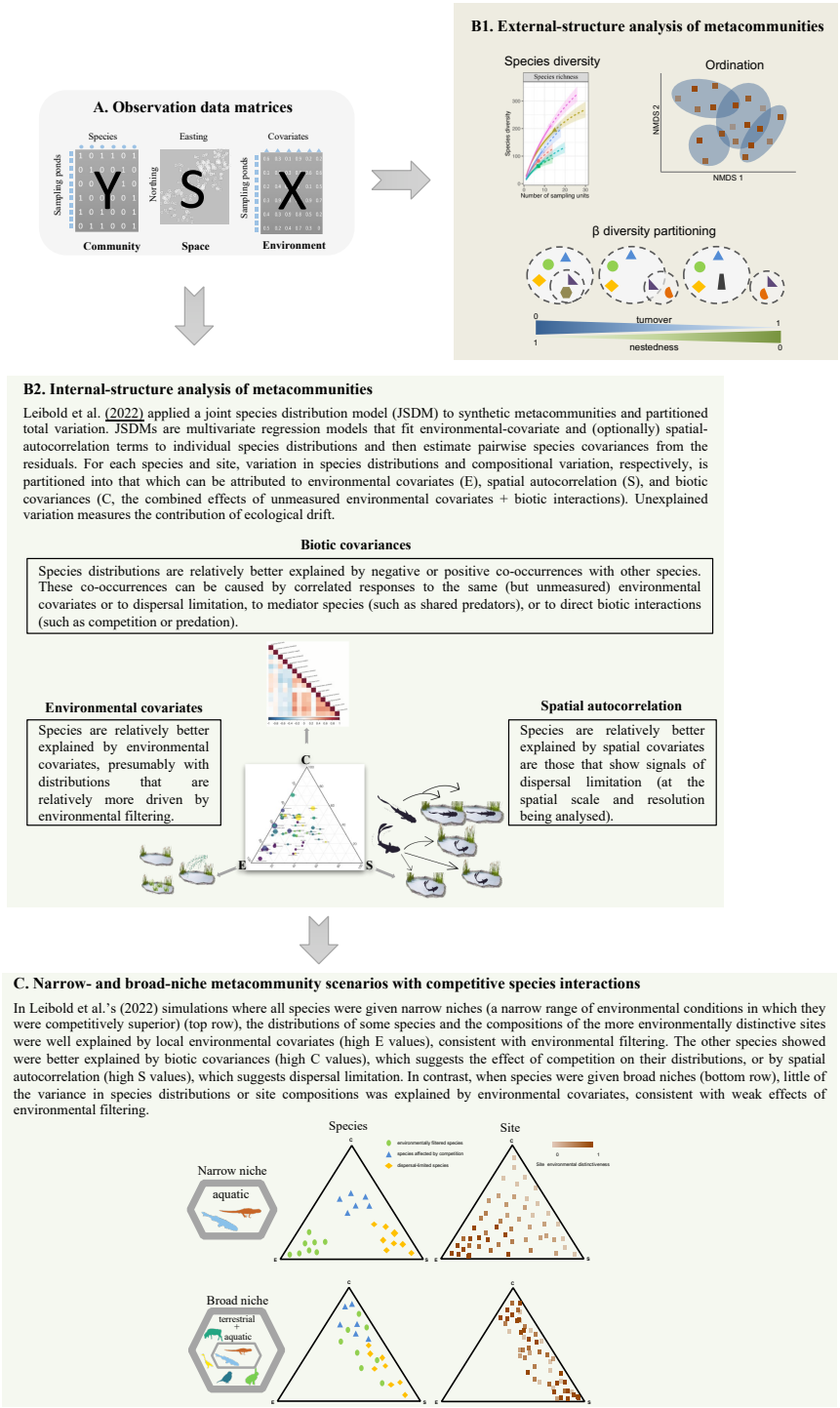
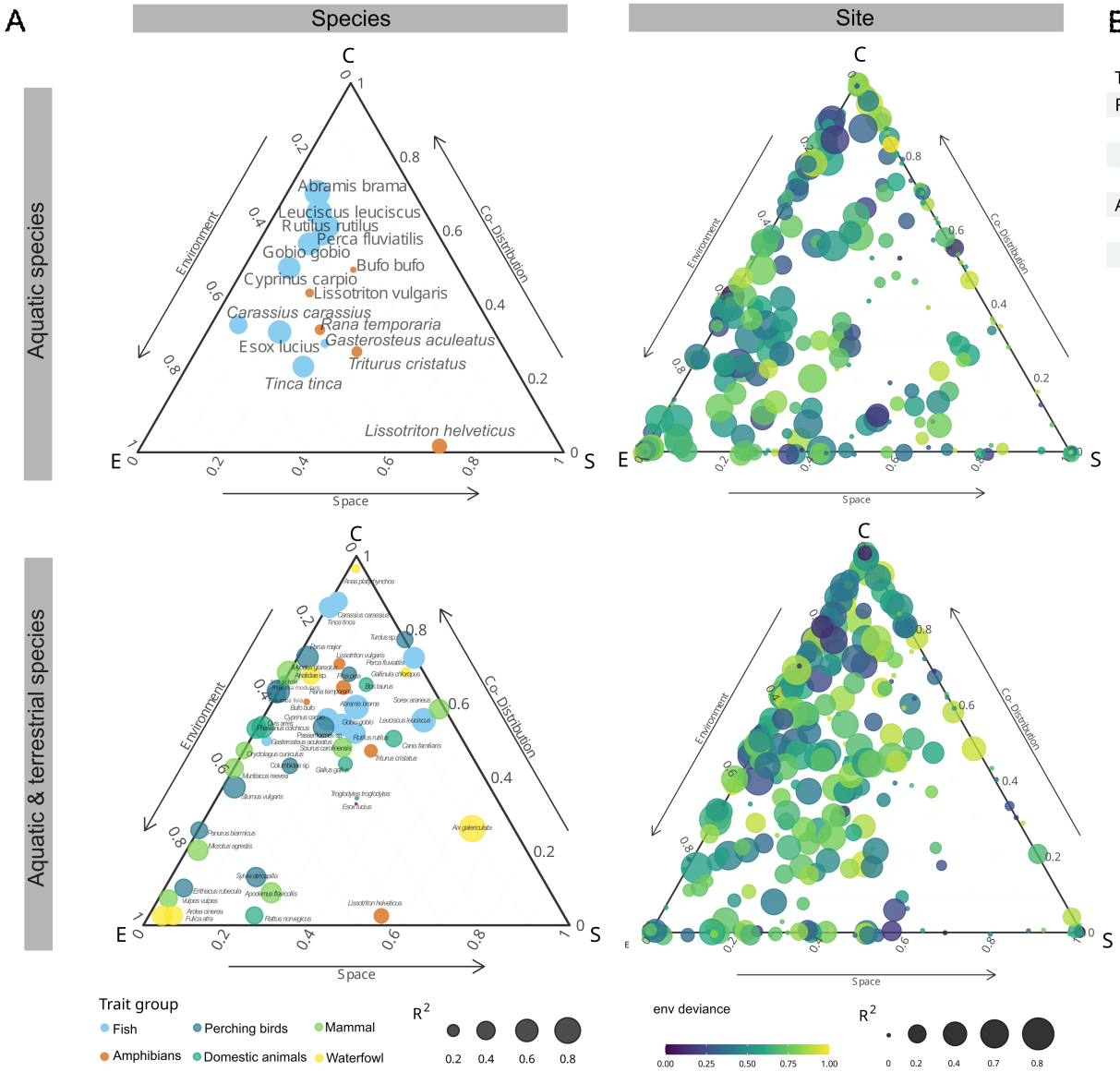


Figure 2. The internal structure of a pond metacommunity. A. The relative contributions of the three components of the internal structure of a pond metacommunity, visualised by decomposing explained variation in species occurrence into the proportions attributable to environmental covariates (E), spatial autocorrelation (S), and biotic covariances (also known as co-distribution) (C). E estimates the relative importance of environmental filtering, S the relative importance of dispersal, and C the joint effect of unmeasured environmental covariates and biotic interactions. E, S and C are plotted together in a ternary plot after dividing each component of variance by its sum to allow comparison among species. Top row. Aquatic species only. Bottom row. Aquatic + Terrestrial species. Left column. Each point is a species, point size scales to total $R^2_{McFadden}$ of each species, and the colours code for trait group. Right column. Each point is a site (pond), point size scales to total $R^2_{McFadden}$ of each site, and the colours code for the site's environmental distinctiveness. The species compositions of more environmentally distinctive ponds are not better explained by environmental covariates. B. Aquatic-species-only and C. Aquatic+terrestrial species. The two tables show the medians and 95% confidence intervals of the E, S, and C values for each trait group. All trait groups have generally low S values.



B

Trait group	median (95% CI)
Fish	
environmental	0.36 (0.23 to 0.59)
biotic covariance	0.53 (0.23 to 0.70)
spatial	0.13 (0.06 to 0.29)
Amphibians	
environmental	0.35 (0.24 to 0.40)
biotic covariance	0.33 (0.01 to 0.50)
spatial	0.26 (0.19 to 0.70)

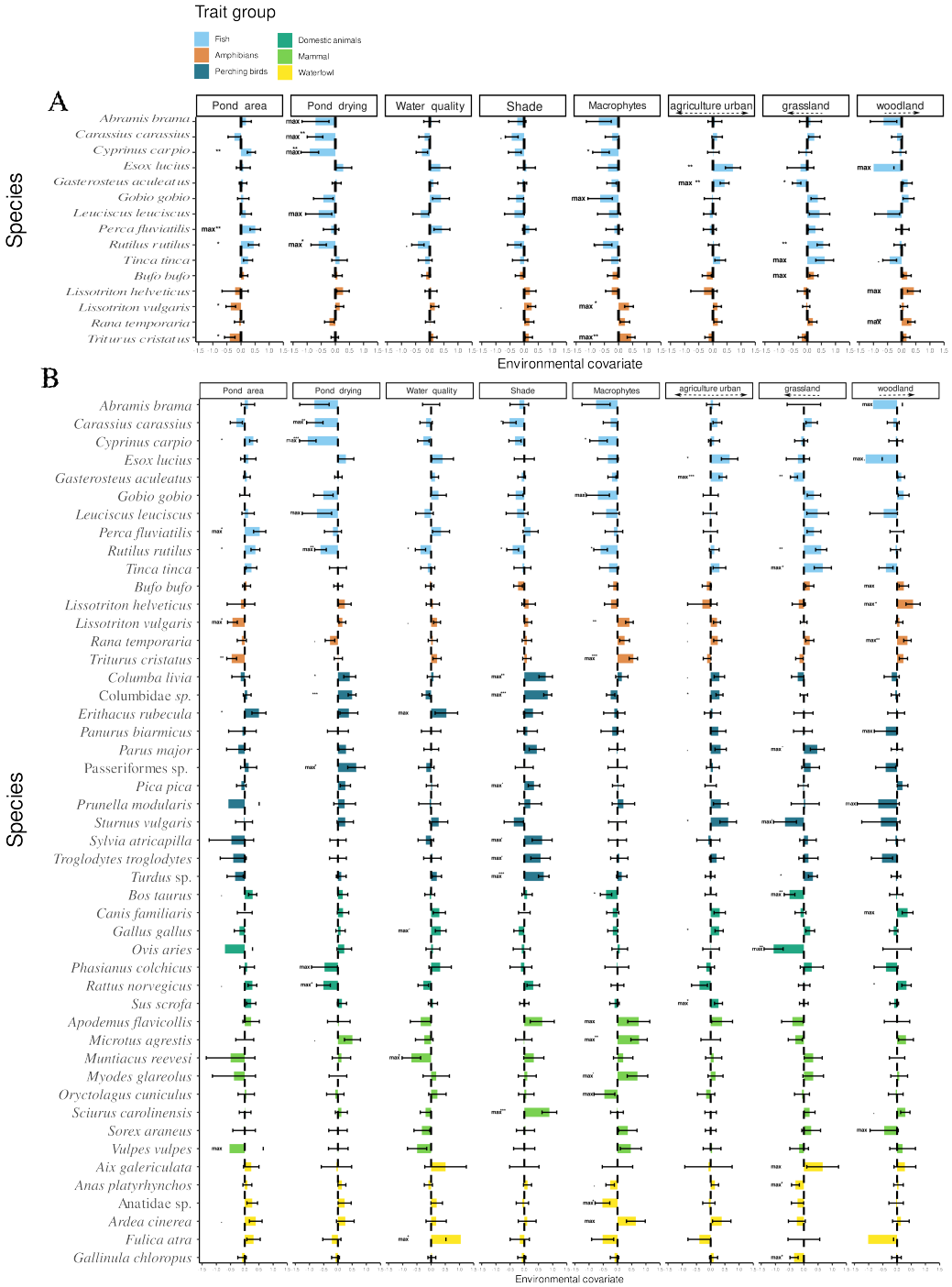
C

Trait group	median (95% CI)
Fish	
environmental	0.21 (0.00 to 0.46)
biotic covariance	0.56 (0.52 to 0.87)
spatial	0.19 (0.01 to 0.38)
Amphibians	
environmental	0.23 (0.19 to 0.44)
biotic covariance	0.60 (0.00 to 0.70)
spatial	0.15 (0.08 to 0.56)
Perching birds	
environmental	0.38 (0.00 to 1.00)
biotic covariance	0.54 (0.00 to 0.77)
spatial	0.03 (0.00 to 0.23)
Domestic animals	
environmental	0.34 (0.15 to 0.74)
biotic covariance	0.53 (0.00 to 0.66)
spatial	0.20 (0.00 to 0.34)
Mammal	
environmental	0.55 (0.00 to 1.00)
biotic covariance	0.45 (0.00 to 0.68)
spatial	0.00 (0.00 to 0.42)
Waterfowl	
environmental	0.18 (0.02 to 1.00)
biotic covariance	0.47 (0.00 to 0.98)
spatial	0.03 (0.00 to 0.64)

652

Figure 3. Estimated environmental effects. Eight environmental covariates were included in the model. The first to fifth covariates from the left are taken from the ten standard pond variables used by surveyors to calculate the Habitat Suitability Index (HSI) of each pond for the great crested newt and are therefore measured at all ponds in our dataset (ARG-UK 2010). The last three covariates describe the dominant land cover class within 500 m of each pond (more details in Table 1). Horizontal bar graphs show the magnitudes, directions, and standard errors of the coefficients of each of the eight environmental covariates for each species. All covariates were normalised before fitting. Significance values are not corrected for multiple comparisons. Colours indicate species groupings.

659



Asterisks (*) indicate environmental covariates that have >95% posterior probability of an effect
Significance codes: P < 0.001: ***; P < 0.01: **; P < 0.05: *; P < 0.1: .

Figure 4. Biotic covariances. A. Aquatic species only, all pairwise covariances. B. Covariances filtered to the 2.5% most negative and positive. C. Aquatic + terrestrial species, all pairwise biotic covariances. D. Covariances filtered to the 2.5% most negative and positive.

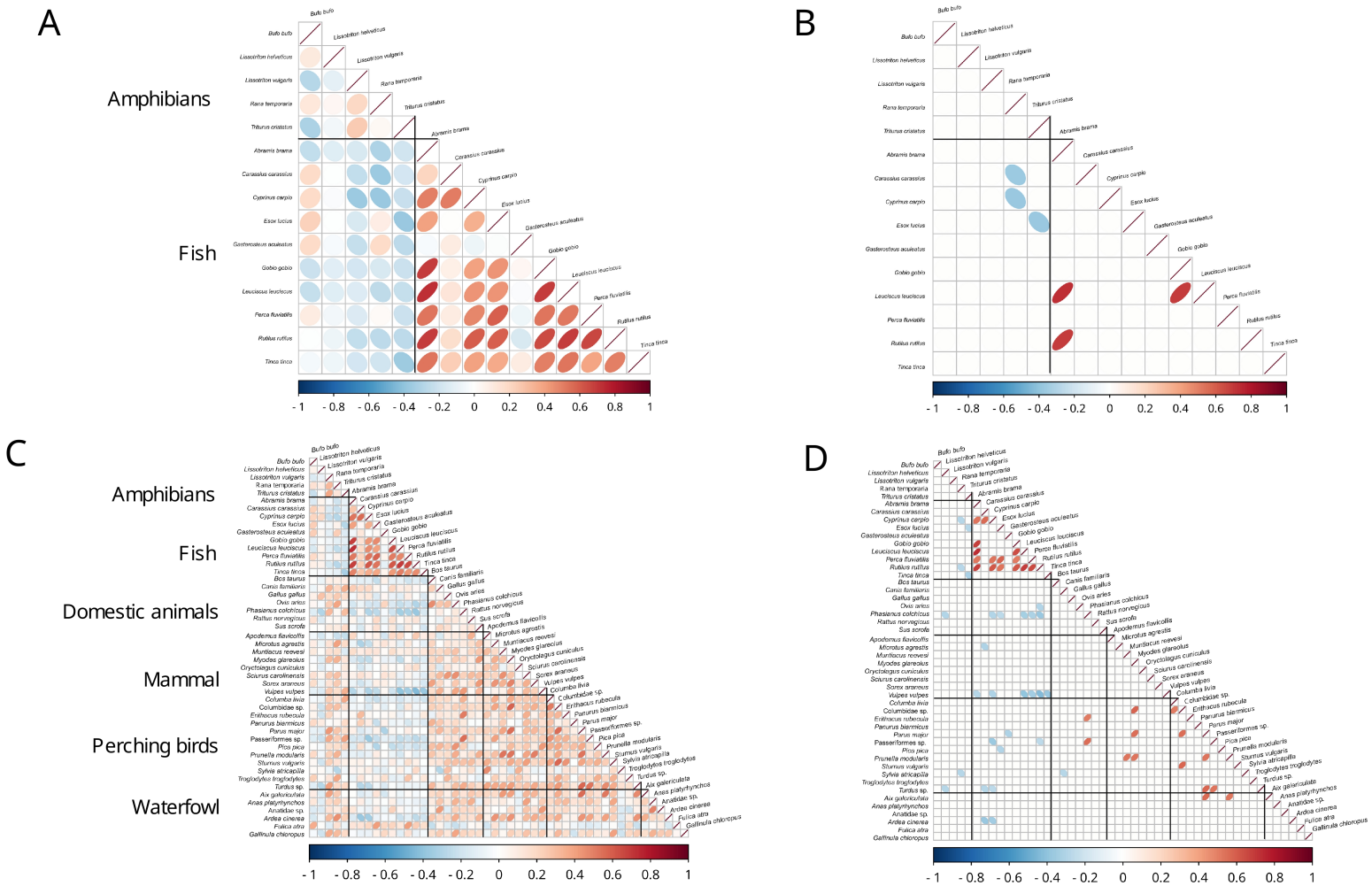
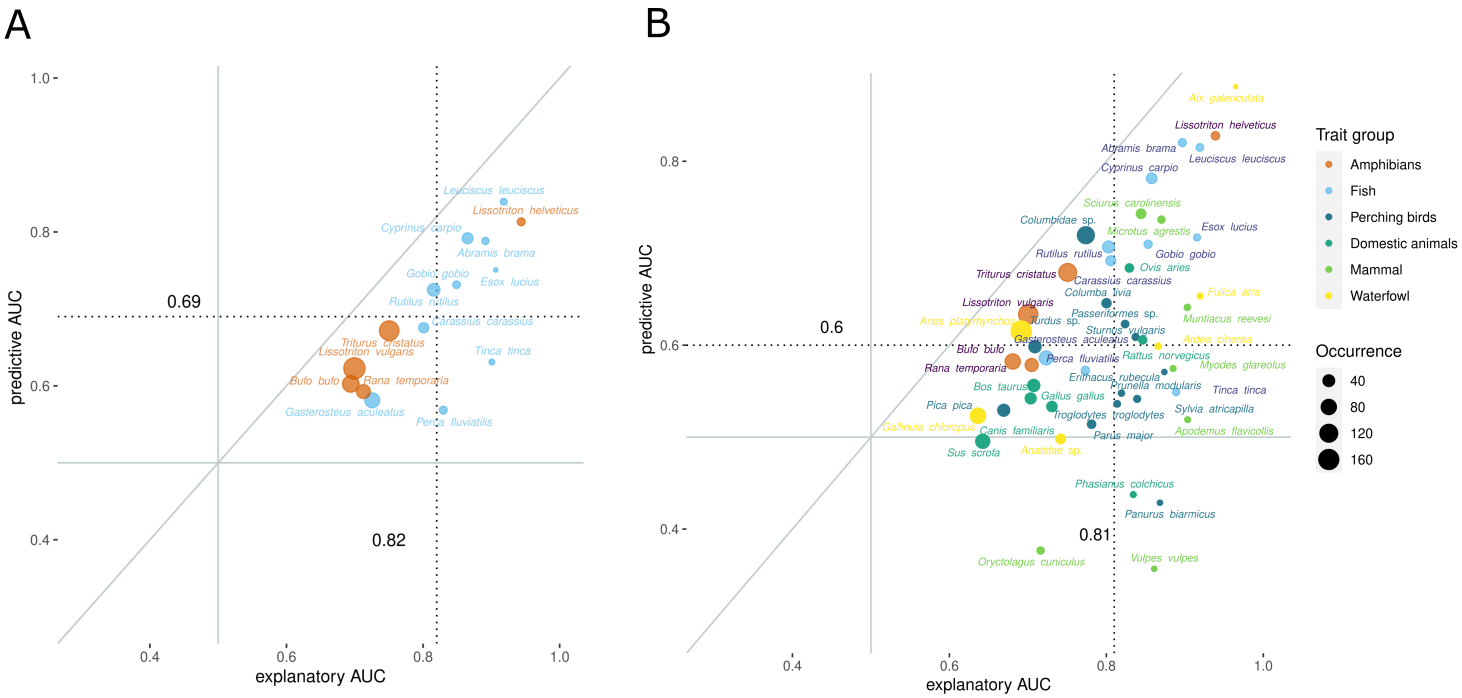


Figure 5. Predictive versus explanatory performance in two JSDMs. A. Aquatic-species-only model. B. Aquatic + Terrestrial species model. Model performance was assessed using the AUC metric (area under the curve), and the dotted lines indicate mean AUC values (vertical for predictive and horizontal for explanatory). Explanatory performance is mostly greater than the predictive performance, as expected when the sample size is adequate. In both models, predictive performance is generally higher for fish than for amphibians. Model overfitting risk is higher for species with lower predictive performance.



668 **Table 1. Environmental covariates included and excluded from the sjSDM models.** The first ten covariates are used to calculate the Habitat
669 Suitability Index (HSI) value of a pond for great crested newt (ARG-UK 2010). We omitted the last five, as they are redundant or potentially
670 redundant with the metabarcoding data (waterfowl, fish scores, pond count), redundant with the land-cover covariates (terrestrial habitat) or lacked
671 variation altogether (geographic location). The last three covariates are the first three principal components extracted from a PCA analysis of land
672 cover within 500 m of each pond, with the land-cover classes extracted from a land-cover map of the UK (Rowland *et al.* 2017).

673

Environmental covariate	Included in model or not	Description
Pond area	Included in model	The surface area of the pond (m^2) when water is at its highest level (excluding flooding events)
Pond permanence	Included in model	How often a pond dries, from 'never dries' to 'dries annually'. Ponds are scored by interviewing the landowner, and if this information is not available, the surveyor makes a judgement based on the water level at the time of the survey. For instance, a pond that is dry by late spring is likely to dry every year, so the pond is described as "Dries annually".
Water quality	Included in model	The rating is based on invertebrate diversity, good water supports an abundant and diverse invertebrate community, poor water quality supports low invertebrate diversity, and bad water quality is clearly polluted.
Shade	Included in model	The percentage of the estimated pond perimeter that is shaded to at least 1m from the bank, usually by trees.
Macrophytes	Included in model	The percentage of the estimated pond surface area occupied by macrophyte cover, including emergent, floating plants (excluding duckweed) and submerged plants reaching the surface.
Waterfowl	Not included	Description of waterfowl density where there is evidence of waterfowl presence and its effect on pond vegetation. We do not include this index because we used metabarcoding to detect waterfowl to species.

Fish	Not included	Presence of fish in the pond, based on local knowledge and surveyor observation. We do not include this in the model because we used metabarcoding to detect fish to species.
Pond count	Not included	Number of ponds within 1 km of the survey pond and not separated by major dispersal barriers such as major roads. As samples were collected on a grid design, one per square kilometre, and surveyors only visit the sampled pond, we did not include this index.
Geographic location	Not included	Because all ponds in this dataset are located in ARG-UK's (2010) Geographical location Zone A, all ponds received the same score, so we do not include it.
Terrestrial habitat	Not included	Subjective surveyor rating of surrounding terrestrial habitat according to perceived suitability for great crested newt. As we are not only estimating great crested newt habitat, we do not include this covariate, but instead use the next three covariates to measure dominant local land cover classes.
Agriculture-urban	Included in model	The first principal component. Positive values indicate increasing arable land cover, and negative values indicate increasing urban land cover. For each pond, the cover of each land-cover class was counted from a satellite-derived land-cover map (Rowland *et al.* 2017). This 21-dimensional dataset was reduced to three principal components (Figure S1).
Grassland	Included in model	The second principal component. Negative values indicate increasing coverage of 'improved grassland'.
Woodland	Included in model	The third principal component. Negative values indicate increasing coverage of woodland.

674

675 **Table 2. Species list and assigned trait group. Incidence is the number of ponds in which each species was found.**

Trait group	Scientific name	Incidence	Trait group	Scientific name	Incidence	Trait group	Scientific name	Incidence
Amphibians	*Lissotriton vulgaris* (smooth newt)	148	Mammal	*Microtus agrestis* (field vole)	11	Perching birds	*Pica pica* (European magpie)	43
Amphibians	*Triturus cristatus* (great crested newt)	121	Mammal	*Oryctolagus cuniculus* (European rabbit)	11	Perching birds	*Columba livia* (pigeon)	21
Amphibians	*Bufo bufo* (common toad)	82	Mammal	*Muntiacus reevesi* (Reeve's muntjac deer)	8	Perching birds	*Parus major* (great tit)	16
Amphibians	*Rana temporaria* (common frog)	54	Mammal	*Apodemus flavicollis* (yellow-necked mouse)	7	Perching birds	Passeriformes sp.	12
Amphibians	*Lissotriton helveticus* (palmate newt)	16	Mammal	*Myodes glareolus* (bank vole)	7	Perching birds	*Sturnus vulgaris* (common starling)	10
Fish	*Gasterosteus aculeatus* (three-spined stickleback)	64	Mammal	*Sorex araneus* (common shrew)	7	Perching birds	*Sylvia atricapilla* (blackcap)	10
Fish	*Rutilus rutilus* (roach)	43	Mammal	*Vulpes vulpes* (red fox)	6	Perching birds	*Troglodytes troglodytes* (wren)	9
Fish	*Cyprinus carpio* (common carp)	31	Domestic species	*Sus scrofa* (pig)	66	Perching birds	*Erithacus rubecula* (European robin)	8
Fish	*Carassius carassius* (crucian carp)	26	Domestic species	*Gallus gallus* (chicken)	45	Perching birds	*Panurus biarmicus* (bearded tit)	6
Fish	*Perca fluviatilis* (perch)	16	Domestic species	*Bos taurus* (cow)	34	Perching birds	*Prunella modularis* (dunnock)	6
Fish	*Gobio gobio* (gudgeon)	15	Domestic species	*Canis familiaris* (dog)	29	Waterfowl	*Anas platyrhynchos* (mallard duck)	167
Fish	*Abramis brama* (common bream)	14	Domestic species	*Ovis aries* (sheep)	16	Waterfowl	*Gallinula chloropus* (common moorhen)	84
Fish	*Leuciscus leuciscus* (dace)	13	Domestic species	*Rattus norvegicus* (brown rat)	15	Waterfowl	Anatidae sp. (duck/geese/swan) sp.	23
Fish	*Tinca tinca* (tench)	11	Domestic species	*Phasianus colchicus* (common pheasant)	7	Waterfowl	*Ardea cinerea* (grey heron)	8
Fish	*Esox lucius* (northern pike)	10	Perching birds	Columbidae sp. (dove) sp.	107	Waterfowl	*Fulica atra* (coot)	7
Mammal	*Sciurus carolinensis* (grey squirrel)	23	Perching birds	*Turdus* sp. (thrush) sp.	46	Waterfowl	*Aix galericulata* (mandarin duck)	5

676