Supporting Information for:

Ground-truthing environmental DNA metabarcoding for ecological hypothesis testing at the pondscape

Lynsey R. Harper[1*], Lori Lawson Handley[1], Christoph Hahn[1,2], Neil Boonham[3,4], Helen C. Rees[5], Erin Lewis[3], Ian P. Adams[3], Peter Brotherton[6], Susanna Phillips[6] and Bernd Hänfling[1]

[1] School of Environmental Sciences, University of Hull, Hull, HU6 7RX, UK
[2] Institute of Zoology, University of Graz, Graz, Styria, Austria
[3] Fera Science Ltd (Fera), Sand Hutton, York, YO14 1LZ, UK
[4] Newcastle University, Newcastle upon Tyne, NE1 7RU, UK
[5] ADAS, School of Veterinary Medicine and Science, The University of Nottingham, Sutton Bonington
[6] Natural England, Peterborough, PE1 1NG, UK

***Corresponding author: Lynsey R. Harper**
School of Environmental Sciences, University of Hull, Hull, HU6 7RX, UK
E-mail: L.Harper@2015.hull.ac.uk

Contents

Appendix 1: Materials and methods

1.1 Samples

In accordance with eDNA sampling methodology outlined by Biggs et al. (2015), 20 x 30 mL water samples were collected at even intervals around the pond margin and pooled in a sterile 1 L Whirl-Pak® stand-up bag, which was shaken to provide a single homogenised sample from each pond. Six 15 mL subsamples were taken from the mixed sample using a sterile plastic pipette (25 mL) and added to sample tubes, containing 33.5 mL absolute ethanol and 1.5 mL sodium acetate 3 M (pH 5.2), for ethanol precipitation. Subsamples were then sent to Fera Science Ltd (Natural England) and ADAS (private contracts) for eDNA analysis according to laboratory protocols established by Biggs et al. (2015). Subsamples were centrifuged at 14,000 x g for 30 minutes at 6 ^{o}C and the supernatant discarded. Subsamples were then pooled during the first step of DNA extraction with the DNeasy Blood & Tissue Kit (Qiagen®, Hilden, Germany), where 360 µL of ATL buffer was added to the first tube, vortexed, and the supernatant transferred to the second tube. This process was repeated for all six tubes. The supernatant in the sixth tube, containing concentrated DNA from all six subsamples, was transferred in a 2 mL tube and extraction continued following manufacturer's instructions to produce one eDNA sample per pond. In 2015, samples were analysed for great crested newt (*Triturus cristatus*) using real-time quantitative PCR (qPCR) and published primers (Thomsen et al. 2012).

1.2 DNA reference database construction

A custom, phylogenetically curated reference database of the target region was created for UK vertebrate species. For freshwater fish, we used a previously created database comprising 67 fish species, which includes all known native and non-native species in the UK and our positive control *Rhamphochromis esox*, a species of cichlid from Lake Malawi (Hänfling et al. 2016). For all remaining vertebrate species recorded in the UK, reference databases were constructed using the ReproPhylo environment (Szitenberg, John, Blaxter, & Lunt, 2015) in a Jupyter notebook (Jupyter Team, 2016). Database curation for each of the main UK vertebrate groups (amphibians, birds, mammals, reptiles) was performed separately to ease data processing. Jupyter notebooks detailing the processing steps for each data subset are deposited in a dedicated GitHub repository for this study (https://github.com/HullUni-bioinformatics/Harper_et_al_2018) which has been permanently archived (https://doi.org/10.5281/zenodo.1304107). Species lists containing the binomial nomenclature of UK vertebrate species were constructed using the Natural History Museum UK Species Database. All vertebrates recorded in the UK were included.

The BioPython script performed a GenBank search based on the species lists and downloaded all available mitochondrial 12S ribosomal RNA (rRNA) sequences for specified species. Where there were no records on GenBank for a UK species, the database was supplemented with downloaded sequences belonging to sister species in the same genus. Species that had no 12S rRNA records on Genbank are provided in Table S1.

Redundant sequences were removed by clustering at 100% similarity using vsearch v1.1 (Rognes, Flouri, Nichols, Quince, & Mahé, 2016). Due to high proportion of partial 12S rRNA records on GenBank for the majority of UK species, only sequences longer than 500 bp were processed initially to increase alignment robustness to large gaps. Sequences were aligned using MUSCLE (Edgar, 2004). Short sequences can cause problems in global paired alignments where the alignment algorithm attempts to align them to longer sequences. Short 12S rRNA sequences (<500 bp) were later incorporated into the existing long 12S rRNA alignment using the hmmer v3 program suite (HMMER development team, 2016) to construct a Hidden Markov Model alignment containing sequences of all lengths. Alignments were trimmed using trimAl (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009). Maximum likelihood trees were inferred with RAxML 8.0.2 (Stamatakis, 2006) using the GTR+gamma model of substitutions. The complete alignments were then processed using SATIVA (Kozlov, Zhang, Yilmaz, Glöckner, & Stamatakis, 2016) for automated identification of 'mislabelled' sequences which could cause conflict in downstream analyses. Putatively mislabelled sequences were removed and process of alignment and phylogenetic tree construction repeated for manual investigation of sequences. The resultant databases (i.e. curated non-redundant reference databases) contained: 198 amphibian sequences from 20/21 species, 112 reptile sequences from 19/20 species, 272 fish sequences from 60/62 species, 940 mammal sequences from 95/112 species, and 622 bird sequences from 347/621 species. Databases for each vertebrate group were concatenated and the combined vertebrate database used for *in silico* validation of primers.

The amphibian database was supplemented by Sanger sequences obtained from tissue of great crested newt, smooth newt (*Lissotriton vulgaris*), Alpine newt (*Mesotriton alpestris*), common toad (*Bufo bufo*), which were supplied by University of Kent under licence from Natural England, and common frog (*Rana temporaria*), supplied by University of Glasgow. Amphibian DNA from University of Kent was extracted from tissue samples using a DNeasy Blood & Tissue kit (Qiagen®, Hilden, Germany) under licence from Natural England by H. Rees. Reference sequences of the entire 12S rRNA region were generated by three sets of novel primers:

Crested newt (61 °C):	Newt_F1	5’-GCACTGAAAATGCTAAGACAGA-3’
	Newt_R6	5’-CAGGTATTTTCTCGGTGTAAGCA-3’
Newts (59 °C):	Newt_F2	5’-GCACTGAAAATGCTAAGACAG-3’
	Newt_R1	5’-TCTCGGTGTAAGCAAGATGC-3’
Anura (57 °C):	AnuraShort_F2	5’-TCCACTGGTCTTAGGAGCCA-3’
	AnuraShort_R1	5’-ACCATGTTACGACTTGCCTC-3’

Primers were designed from an alignment of tRNA, 12S and 16S rRNA regions in UK Caudata and Anura species. PCR reactions were performed in 25 µL volumes containing: 12.5 µL of MyTaq™ Red Mix (Bioline Reagents Limited, London, UK), 1 µL (final concentration - 0.04 µM) of forward and reverse primer (Integrated DNA Technologies, Belgium), 8.5µL of molecular grade sterile water (Fisher Scientific UK Ltd, Loughborough, UK) and 2 µL DNA template. PCRs were performed on an Applied Biosystems® Veriti Thermal Cycler (Fisher Scientific UK Ltd, Loughborough, UK) with the following profile: 95 °C for 3 min, 35 cycles of 95 °C for 30 sec, x °C (see temperatures above) for 60 sec and 72 °C for 90 sec, followed by a final elongation step at 72 °C for 10 min. Purified PCR products were Sanger sequenced directly (Macrogen Europe, Amsterdam, Netherlands) in both directions using the PCR primers. Sequences were edited using CodonCode Aligner (CodonCode Corporation, Centerville, MA, USA). The complete reference database compiled in GenBank format has been deposited in the GitHub repository for this study.

1.3 Primer validation

Vertebrate DNA from eDNA samples was amplified with published 12S rRNA primers 12S-V5-F (5’-ACTGGGATTAGATACCCC-3’) and 12S-V5-R (5’-TAGAACAGGCTCCTCTAG-3’) (Riaz et al., 2011). Primers were validated for the present study *in silico* using ecoPCR software (Ficetola et al., 2010) against a custom, phylogenetically curated reference database for UK vertebrates. Parameters were set to allow a fragment size of 50-250 bp and maximum of three mismatches between the primer pair and each sequence in the reference database. Primers were previously validated *in vitro* for UK fish communities by Hänfling et al. (2016) and here were also validated against tissue DNA extracted from UK amphibian species: great crested newt, smooth newt, palmate newt (*Lissotrition helveticus*), Alpine newt, common frog and common toad. Primer validation tests were performed at University of Hull in a separate laboratory situated on a different floor to the dedicated eDNA laboratory. A dilution series (10^0 to 10^{-8}) was performed for DNA (standardised to 5 ng/µL) from each species to identify the limit of detection (LOD) for each species. Molecular grade sterile water (Fisher Scientific UK Ltd, Loughborough, UK) substituted template DNA for the PCR negative control.

1.4 eDNA metabarcoding

A two-step PCR protocol was performed on eDNA samples at University of Hull. Dedicated rooms were available for pre-PCR and post-PCR processes. Pre-PCR processes were performed in a dedicated eDNA laboratory, with separate rooms for filtration, DNA extraction and PCR preparation of sensitive environmental samples. PCR reactions were set up in a UV and bleach sterilized laminar flow hood. Eight-strip PCR tubes with individually attached lids were used instead of 96-well plates to minimise cross-contamination risk between samples (Port et al., 2016). After the first sequencing run revealed substantial human contamination across samples and PCR controls, reactions prepared for the second sequencing run were sealed with mineral oil as an additional measure against PCR contamination. For the first PCR, three replicates were performed for each sample to combat PCR stochasticity. Alternating PCR positive and negative controls were included on each PCR strip (six positive and negative controls on each 96-well plate), to screen for sources of potential contamination. The DNA used for the PCR positive control was *R. esox,* as occurrence in UK ponds is extremely rare or non-existent. The negative control substituted molecular grade sterile water (Fisher Scientific UK Ltd, Loughborough, UK) for template DNA.

During the first PCR, the target region was amplified using the primers described above, including adapters (Illumina, 2011). First step PCR reactions were performed in a final volume of 21.1 µL, using 2 µL of DNA extract as a template. The amplification mixture contained 10.5 µL of MyTaq™ HS Red Mix (Bioline Reagents Limited, London, UK), 1.05 µL (final concentration - 0.5 µM) of forward and reverse primer (Integrated DNA Technologies, Belgium) and 6.5 µL of molecular grade sterile water (Fisher Scientific UK Ltd, Loughborough, UK). PCR was performed on an Applied Biosystems® Veriti Thermal Cycler (Fisher Scientific UK Ltd, Loughborough, UK) and PCR conditions for the first component of the two-step protocol consisted of: an incubation step at 98 °C for 5 min, followed by 35 cycles of denaturation at 98 °C for 15 s, annealing at 56 °C for 20 s, and extension at 72 °C for 30 s with final extension at 72 °C for 10 min. PCR products were stored at 4 °C until fragment size was verified by visualising 5 µL of selected PCR products on 2% agarose gels (100 mL 0.5x TBE buffer, 2 g agarose powder). Gels were then stained with ethidium bromide and imaged using Image Lab Software (Bio-Rad Laboratories Ltd, Watford, UK). A PCR product was deemed positive where there was an amplification band on the gel that was of the expected size (200-300 bp). PCR replicates for each sample were pooled in preparation for the addition of Illumina indexes in the second PCR, which resulted in 63.3 µL of PCR product for each sample. PCR positive and negative controls were not pooled to allow individual purification and sequencing of all 228 PCR controls. All PCR products (30 µL samples and 15 µL PCR controls) were then purified to remove excess primer using E.Z.N.A.® Cycle Pure V-Spin Clean-Up Kits (Omega Bio-tek, GA, USA) following manufacturers protocol. Eluted DNA was stored at -20 °C until the second PCR could be performed.

In the second PCR, Multiplex Identification (MID) tags (unique 8-nucleotide sequences) and Illumina MiSeq adapter sequences were bound to the amplified product. These tags were included in the forward and reverse primers resulting in indexed primers for second PCR (O'Donnell, Kelly, Lowell, & Port, 2016). For each second PCR plate, 96 unique tag combinations were created by combining eight unique forward tags with 12 unique reverse tags or vice versa (Kitson et al., 2018). A total of 384 unique tag combinations were achieved, allowing samples to be distinguished during bioinformatics analysis. Second step PCR reactions were performed in eight-strip PCR tubes with individually attached lids in a final volume of 21.1 µL, using 2 µL of purified DNA from the first PCR product as a template. The amplification mixture contained 10.5 µL of MyTaq™ HS Red Mix (Bioline Reagents Limited, London, UK), 2.1 µL (final concentration - 0.5 µM) of tagged primer mix (Integrated DNA Technologies, Belgium) and 6.5 µL of molecular grade sterile water (Fisher Scientific UK Ltd, Loughborough, UK). PCR was performed on an Applied Biosystems® Veriti Thermal Cycler (Fisher Scientific UK Ltd, Loughborough, UK) with the following profile: denaturation at 95 °C for 3 min, followed by 12 cycles of annealing at 98 °C for 20 s and extension at 72 °C for 30 s with final extension at 72 °C for 5 min. PCR products were stored at 4 °C before they were all visualised on 2% agarose gels (100 mL 0.5x TBE buffer, 2 g agarose powder) using 5 µL PCR product. Gels were then stained with ethidium bromide and imaged using Image Lab Software (Bio-Rad Laboratories Ltd, Watford, UK). Again, PCR products were deemed positive where there was an amplification band on the gel that was of the expected size (200-300 bp). Amplification bands were found to be present in some of the negative controls thus all negative controls were included for sequencing.

All remaining library preparation was conducted at Fera Science Ltd. PCR products were transferred to a new 96-well PCR plate for individual purification with AMPure® XP beads (Beckman Coulter (UK) Ltd, High Wycombe, UK) and an invitrogen® magnetic stand (Fisher Scientific UK Ltd, Loughborough, UK). The Illumina PCR clean-up protocol was adapted to use 18.6 µL AMPure® XP beads (1.2x PCR product) to 15-16 µL PCR product. Illumina protocol was then followed until the beads were resuspended in 15 µL molecular grade water and incubated at room temperature for 5 minutes. The supernatant without beads in each well were not transferred to a new plate due to low volumes of purified product. Further pipetting may have resulted in loss of DNA. Each plate was sealed and stored at 4 °C until quality assurance. An Invitrogen™ Quant-IT™ PicoGreen™ dsDNA Assay (Fisher Scientific UK Ltd, Loughborough, UK) was conducted for all samples on a Fluoroskan™ Microplate Fluorometer (Life Technologies Ltd, Paisley, UK). Samples were then normalised and pooled to create 4 nM pooled libraries before quantification using an Invitrogen™ Qubit™ dsDNA HS Assay Kit (Fisher Scientific UK Ltd, Loughborough, UK). Both libraries passed quality assurance with concentrations of 2.62 ng/µl and 4.14 ng/µl respectively. An Agilent 4200 Tapestation System (Agilent Technologies, Santa Clara, CA, United States) was then used to check and compare size of the pooled libraries to selected samples. The pooled libraries were 272 bp and 299 bp (expected 286 bp) with samples in the same range. Equimolar libraries (4 nM) were then created using tapestation trace size estimates and Qubit concentrations. Libraries were run at 12 pM concentration on an Illumina MiSeq using 2 x 300 bp V3 chemistry (Illumina Inc., San Diego, CA, USA). Both libraries included a 10% PhiX DNA spike-in control to improve clustering during initial sequencing.

Illumina data was converted from raw sequences to taxonomic assignment using a custom pipeline for reproducible analysis of metabarcoding data: metaBEAT (metaBarcoding and eDNA Analysis Tool) v0.8 (https://github.com/HullUni-bioinformatics/metaBEAT). Bioinformatic analysis using metaBEAT largely followed the workflow outlined by Hänfling et al. (2016) for sample processing and taxonomic assignment of sequenced eDNA samples from Windermere. Adaptations to this workflow are described (see also Harper et al. 2018): raw reads were quality trimmed using Trimmomatic v0.32 (Bolger, Lohse, & Usadel, 2014), both from the read ends (minimum per base phred score Q30), as well as across sliding windows (window size 5bp; minimum average phred score Q30). Reads were clipped to a maximum length of 110 bp and reads shorter than 90 bp after quality trimming were discarded. To reliably exclude adapters and PCR primers, the first 25 bp of all remaining reads were also removed. Sequence pairs were merged into single high quality reads using FLASH v1.2.11 (Magoč & Salzberg, 2011), if a minimum of 10 bp overlap with a maximum of 10% mismatch was detected between pairs. For reads that were not successfully merged, only forward reads were kept. To reflect our expectations with respect to fragment size, a final length filter was applied and only sequences of length 80-120 bp were retained. These were screened for chimeric sequences against our custom reference database using the uchime algorithm (Edgar, Haas, Clemente, Quince, & Knight, 2011), as implemented in vsearch v1.1 (Rognes et al., 2016). Redundant sequences were removed by clustering at 97% identity ('--cluster_fast' option) in vsearch v1.1 (Rognes et al., 2016). Clusters represented by less than five sequences were considered sequencing error and omitted from further analyses. Non-redundant sets of query sequences were then compared against our custom reference database using BLAST (Zhang, Schwartz, Wagner, & Miller, 2000). For any query matching with at least 98% identity to a reference sequence across more than 80% of its length, putative taxonomic identity was assigned using a lowest common ancestor (LCA) approach based on the top 10% BLAST matches. Sequences that could not be assigned (non-target sequences) were subjected to a separate BLAST search against the complete NCBI nucleotide (nt) database at 98% identity to determine the source via LCA as described above. To ensure reproducibility of analyses, the described workflow has been deposited in the GitHub repository.

1.5 Data Analysis

A supplementary analysis was performed where a series of blanket false positive sequence thresholds (0.05 - 30%) were applied to the dataset to ensure results did not differ drastically from species-specific thresholds (see Tables S4-9).

1.5.1 Individual species associations

Species associations between all vertebrates were investigated using presence-absence data generated by eDNA metabarcoding with the method of Veech (2013) implemented in the R package ‘cooccur’ v1.3 (Griffith, Veech, & Marsh, 2016). This is a probabilistic model which measures species co-occurrence (presence-absence) as the number of sampling sites where two species co-occur. The observed co-occurrence of a given dataset is compared to the expected co-occurrence. Expected co-occurrence is determined by the probabilities of each species’ occurrence multiplied by the number of sampling sites. Effect sizes were also computed for species pairs to examine species associations regardless of statistical significance. These are equivalent to the difference between expected and observed frequency of co-occurrence. The values are then standardized by dividing these differences by sample size. In standardized form, these values are bounded from -1 to 1, with positive values indicating positive associations and negative values indicating negative associations.

1.5.2 Biotic and abiotic determinants of great crested newt occurrence

Collinearity and spatial autocorrelation within the dataset were investigated before the most appropriate regression model was determined. Collinearity between explanatory variables was assessed using a Spearman's rank pairwise correlation matrix. Collinearity was observed between pond circumference, pond length, pond width, and pond area. Pond area encompasses length and width thus taking the same measurements and accounting for the same variance in the data as these variables. Therefore, pond circumference, pond length, and pond width were removed from the dataset so as remaining variables were not highly correlated (Zuur, Ieno, Walker, Saveliev, & Smith, 2009). Shading (percentage of total pond margin shaded) and terrestrial overhang (percentage of pond overhung by trees and shrubs) were also collinear. As terrestrial overhang accounts for shading of the entire pond, whereas shading considers only the pond margin, terrestrial overhang was retained as an explanatory variable. After collinear variables were removed, variance inflation factors (VIFs) of remaining variables were calculated using the

R package 'car' v2.1-6 (Fox & Weisberg, 2011) to identify remnant multicollinearity. Multicollinearity (VIF > 3) (Zuur et al., 2009) was still present in Habitat Suitability Index (HSI) score and HSI band. Many of the environmental variables are also used as indices to calculate HSI score thus HSI score may mask variation caused by these variables individually. HSI score and HSI band were removed prior to model selection.

A large number of explanatory variables remained: max. depth; area; density, overhang; macrophyte cover; permanence; water quality; pond substrate; inflow; outflow; pollution; presence of amphibians, waterfowl and fish; woodland; rough grass; scrub/hedge; ruderals; terrestrial other; and overall terrestrial habitat quality. The relative importance of these for determining great crested newt occurrence was inferred using a classification tree within the R package 'rpart' v4.1-13 (Therneau, Atkinson, & Ripley, 2014). The classification tree suggested the most important explanatory variables of great crested newt occurrence were: smooth newt presence, species richness, maximum depth of ponds, fish presence, pond density, pond area, amphibian presence, waterfowl presence (which incorporates identified species associations between great crested newt and common moorhen, *Gallinula chloropus*, and Eurasian coot, *Fulica atra*), terrestrial habitat, pond substrate, grey squirrel (*Sciurus carolinensis*) presence, three-spined stickleback presence (*Gasterosteus aculeatus*), pond outflow, macrophyte cover, water quality and pond permanence. Smooth newt, grey squirrel and three-spined stickleback were also identified as having significant associations with great crested newt by the co-occurrence analysis. A pruning diagram was applied to the data to cross-validate the classification tree and remove unimportant explanatory variables. A tree of six was optimal according to the pruning diagram, indicating that six explanatory variables should be retained for statistical analysis. Many variables occurred more than once in the classification tree, indicative of weak non-linear relationships with the response variable. Generalised Additive Models (GAMs) were performed to deal with non-linearity but several explanatory variables were in fact linear (estimated one degree of freedom for smoother) (Zuur et al., 2009).

The ponds in this study had restricted spatial distribution and were nested within three UK counties (Figure S1) thus spatial autocorrelation may be present. This phenomena is common in ecological studies of species presence-absence as sites located within an animal's ranging capability are likely to be inhabited (Zuur et al., 2009). Great crested newt individuals can migrate distances of 1-2 km to new ponds (Edgar & Bird, 2006; Haubrock & Altrichter, 2016), thus occurrence of great crested newt is likely in ponds that are closely located to one another in a given area. Spline correlograms - graphical representations of spatial correlation between locations at a range of lag distances that are smoothed using a spline function (Bjørnstad, 2009) - were constructed using R package 'ncf' v1.1-7 to examine spatial autocorrelation between ponds. Spline correlograms of the pearson residuals of the raw data, a binomial Generalised Linear Model (GLM), and a binomial Generalised Linear Mixed Model (GLMM) were compared. GLMMs can account for dependencies within sites, handled with the introduction of random effects (Zuur et al., 2009). Each eDNA sample represented a different pond and thus sample was treated as a random effect. The GLMM successfully accounted for spatial dependencies between ponds based on the spline correlogram of the Pearson residuals.

A series of alternative mixed effects models that covered different combinations of explanatory variables to test different hypotheses were then evaluated. Explanatory variables were grouped into functional groups. For example, pond properties, terrestrial habitat and pond biodiversity. The GLMM containing only presence of species or guilds had the lowest AIC value but as we were also interested in habitat predictors of great crested newt, model selection was performed on the GLMM containing all explanatory variables.

1.5.3 Biotic and abiotic determinants of vertebrate species richness

The species richness classification tree indicated that terrestrial overhang was the most important explanatory variable, followed by amphibian presence, rough grass habitat, pond density, maximum pond depth, pond area, woodland, ruderals, pollution, fish presence, terrestrial other, macrophyte cover, pond outflow, water quality, waterfowl presence, pond inflow, scrub/hedge and pond permanence. A tree of three or five was optimal according to the pruning diagram, indicating that three or five explanatory variables should be retained for statistical analysis.

Appendix 2: Results

2.1 Primer validation

The *in silico* analysis confirmed high taxonomic coverage (59.0% of target vertebrate species amplified) and resolution of the 12S rRNA primers. A wide range of UK vertebrate taxa were amplified, with fragment length ranging from 90-114 bp. The primers amplified 16/21 amphibian species, including great crested newt. Palmate newt, Italian crested newt (*Triturus carnifex*), brown cave salamander (*Hydromantes genei*)*,* marsh frog (*Pelophylax esculentus*) and agile frog (*Rana dalmatina*) were not amplified *in silico*. All sequences from these species were manually aligned to the primers using the alignment viewer and editor AliView (Larsson, 2014), confirming potential for amplification. The primers amplified 47/67 fish species, including the threatened European eel (*Anguilla anguilla*), but amplification of UK freshwater fish assemblages was confirmed *in vitro* by Hänfling et al. (2016). The primers amplified 14/20 reptile species including slow worm (*Anguis fragilis*) and common lizard (*Zootoca vivipara*). Reference sequences were not available for one species and a further five species were not amplified. Primers were only validated for 282/621 bird species (including common waterfowl species). There were no 12S rRNA data available for 243/621 bird species and a further 96 species were not amplified. Similarly, no reference data were available for nine mammal species (bats and marine mammals) and a further 15 species were not amplified. Only 88/112 mammal species were validated. Several marine mammal species were not amplified but would not be found in freshwater ponds. However, priority species for freshwater management, such as water vole *Arvicola amphibius* and American mink *Mustela vison*, were not amplified alongside other species of bat, vole and shrew that may frequent ponds. During *in vitro* tests, bands were observed by agarose gel electrophoresis for all amphibian tissue tested, including palmate newt which was not amplified *in silico*, and no bands were observed in NTCs. The LOD was variable for each species: great crested newt, palmate newt, common frog and common toad were not amplified below 5 x 10^{-4} ng/µl, whereas Alpine newt was not amplified below 5 x 10^{-3} ng/µl and smooth newt below 5 x 10^{-5} ng/µl. Due to sheer number of and legislation surrounding many UK amphibian, reptile, bird and mammal species, *in vitro* testing for all target taxa was unfeasible and metabarcoding proceeded on the basis of *in silico* amplification.

2.2 Biotic and abiotic determinants of great crested newt occurrence

The co-occurrence analysis revealed of 1770 species pair combinations. 1406 pairs (79.44%) were removed from the analysis because expected co-occurrence was less than one, leaving 364 pairs for analysis. The pairwise combinations revealed 17 negative and 48 positive significant co-occurrence patterns. The remaining co-occurrence patterns were random thus the observed presence-absence data did not significantly deviate from the expected presence-absence data. No pairs were unclassifiable indicative of sufficient statistical power to analyse all pairs. A pairing profile was constructed to understand each species' individual contribution to the positive and negative species associations. Interactions were clustered in a few species rather than being evenly distributed. When observed and expected co-occurrence was examined, some species pairs deviated from the expected co-occurrence. A minority of species pairs exhibited fewer than expected co-occurrences but these pairs were largely clustered towards having low expected co-occurrence.

Appendix 3: Tables

Table S1 Summary of environmental metadata on pond characteristics and surrounding terrestrial habitat included in analysis of crested newt occupancy and vertebrate species richness.

Variable	Description	Unit/categories
Maximum depth	Depth of pond	m
Circumference	Pond circumference	m
Width	Pond width	m
Length	Pond length	m
Area	Pond area	m^2
Density	Pond density	Number of ponds per km^2
Terrestrial overhang	Percentage of pond overhung by trees and shrubs	%
Shading	Percentage of total pond margin shaded to at least 1 m from the shore	%
Macrophyte cover	Percentage of pond surface occupied by macrophytes	%
Habitat Suitability Index (HSI)	Score calculated from aforementioned variables which indicates habitat quality for crested newt (0 = poor, 1 = excellent)	Decimal
Habitat Suitability Index (HSI) band	Categorical classification of HSI score	Poor/below average/average/good
Pond permanence	Pond permanence	Dries annually/rarely dries/sometimes dries/ never dries
Water quality	Subjective assessment based on invertebrate diversity, presence of submerged vegetation, and knowledge of water inputs to pond.	Bad/poor/moderate/good/excellent
Pond substrate	Type of substrate	Not known/rock/clay/concrete/sand, gravel, pebbles/lined/peat-organic
Inflow	Water inputs to pond	Absent/present
Outflow	Water leaving pond	Absent/present

Pollution	Rubbish or other signs of pollution	Absent/present
Other amphibians	Presence of amphibian species other than crested newt	Absent/present
Fish	Presence of any fish species	Absent/possible/minor/major
Waterfowl	Presence of any waterfowl species	Absent/minor/major
Woodland	Terrestrial habitat: woodland	None/some/important
Rough grass	Terrestrial habitat: rough grass	None/some/important
Scrub/hedge	Terrestrial habitat: scrub/hedge	None/some/important
Ruderals	Terrestrial habitat: ruderals	None/some/important
Terrestrial other	Other good quality terrestrial habitat that does not conform to aforementioned habitat types	None/some/important
Overall terrestrial habitat score	Overall quality of terrestrial habitat	None/poor/moderate/good

Table S2 List of species for which no 12S rRNA records were available on Genbank. Only UK species which had no records for sister species within the same genus are included.

Common name	**Binomial nomenclature**
North Atlantic right whale	*Eubalaena glacialis*
Common kingfisher	*Alcedo atthis*
Trumpeter finch	*Bucanetes githagineus*
Green heron	*Butorides virescens*
Greater short-toed lark	*Calandrella brachydactyla*
Lesser short-toed lark	*Calandrella rufescens*
Lapland longspur	*Calcarius lapponicus*
Wilson’s warbler	*Cardellina pusilla*
Rufuous-tailed scrub robin	*Cercotrichas galactotes*
MacQueen’s bustard	*Chlamydotis macqueenii*
Lark sparrow	*Chondestes grammacus*
White-throated dipper	*Cinclus cinclus*
Great spotted cuckoo	*Clamator glandarius*
Long-tailed duck	*Clangula hyemalis*
Corn crake	*Crex crex*
Crested lark	*Galerida cristata*
European storm petrel	*Hydrobates pelagicus*
Little gull	*Hydrocoloeus minutus*
White-throated robin	*Irania gutturalis*
Hooded merganser	*Lophodytes cucullatus*
European crested tit	*Lophophanes cristatus*
Woodlark	*Lullula arborea*
Siberian blue robin	*Larvivora cyane*
Rufous-tailed robin	*Larvivora sibilans*
Thrush nightingale	*Luscinia luscinia*
Common nightingale	*Luscinia megarhynchos*
Bluethroat	*Luscinia svecica*
Black scoter	*Melanitta americana*
Velvet scoter	*Melanitta fusca*

Common scoter	*Melanitta nigra*
Surf scoter	*Melanitta perspicillata*
Bimaculated lark	*Melanocorypha bimaculata*
Calandra lark	*Melanocorypha calandra*
White-winged lark	*Melanocorypha leucoptera*
Black lark	*Melanocorypha yeltoniensis*
Song sparrow	*Melospiza melodia*
Black-and-white warbler	*Mniotilta varia*
Common rock thrush	*Monticola saxatilis*
Blue rock thrush	*Monticola solitarius*
Wilson’s storm petrel	*Oceanites oceanicus*
Band-rumped storm petrel	*Oceanodroma castro*
Leach’s storm petrel	*Oceanodroma leucorhoa*
Swinhoe’s storm petrel	*Oceanodroma monorhis*
Tennessee warbler	*Oreothlypis peregrina*
Northern waterthrush	*Parkesia noveboracensis*
Savannah sparrow	*Passerculus sandwichensis*
Rosy starling	*Pastor roseus*
American cliff swallow	*Petrochelidon pyrrhonota*
Steller’s eider	*Polysticta stelleri*
Eurasian crag martin	*Ptyonoprogne rupestris*
Sand martin	*Riparia riparia*
Whinchat	*Saxicola rubetra*
African stonechat	*Saxicola torquatus*
Northern parula	*Setophaga americana*
Hooded warbler	*Setophaga citrina*
American yellow warbler	*Setophaga petechia*
American redstart	*Setophaga ruticilla*
Wallcreeper	*Tichodroma muraria*
Brown thrasher	*Toxostoma rufum*
Golden-winged warbler	*Vermivora chrysoptera*

Table S3 List of species detected in PCR positive controls by eDNA metabarcoding and corresponding species-specific false positive sequence threshold applied.

Common name	**Binomial name**	**Species-specific false positive sequence threshold**
European eel	*Anguilla anguilla*	0.000094
Common carp	*Cyprinus carpio*	0.000163
Common minnow	*Phoxinus phoxinus*	0.001287
Common roach	*Rutilus rutilus*	0.000291
European chub	*Squalius cephalus*	0.004080
Three-spined stickleback	*Gasterosteus aculeatus*	0.066667
Atlantic herring	*Clupea harengus*	0.000115
Common toad	*Bufo bufo*	0.066667
Common frog	*Rana temporaria*	0.000596
Smooth newt	*Lissotriton vulgaris*	0.066667
Great crested newt	*Triturus cristatus*	0.000276
Green-winged teal	*Anas carolinensis*	0.000322
Eurasian coot	*Fulica atra*	0.000223
Common moorhen	*Gallinula chloropus*	0.000179
Common starling	*Sturnus vulgaris*	0.000139
Human	*Homo sapiens*	0.253333
Brown rat	*Rattus norvegicus*	0.000467
Cow	*Bos taurus*	0.003542
Pig	*Sus scrofa*	0.000877

Table S4 Effect of number of species in different vertebrate groups on great crested newt occupancy as determined using a binomial GLMM for different metabarcoding sequence thresholds (*N* = 532 ponds). For categorical variables with more than one level, effect size and standard error are only given for levels reported in the model summary. Test statistic is for LRT used. Significant P-values (<0.05) are in bold.

Threshold	**Overdispersion**	**Model fit**	**Model variables**	**Effect size**	**Standard error**	χ^2	***P***
No threshold	χ^2_{525} = 519.016	χ^2_{8} = 18.319	**Fish**	-0.215	0.101	4.913	**0.027**
	P = 0.566	*P* = 0.019	**Amphibian**	0.454	0.120	16.528	**<0.001**
		R^2 = 10.10%	**Waterfowl**	0.523	0.163	11.070	**0.001**
			Terrestrial bird	-0.435	0.277	2.715	0.099
			Mammal	0.146	0.082	3.224	0.073
0.05%	χ^2_{525} = 526.993	χ^2_{8} = 56.79	**Fish**	-0.238	0.121	4.224	**0.040**
	P = 0.467	*P* < 0.001	**Amphibian**	0.338	0.127	7.723	**0.006**
		R^2 = 6.93%	**Waterfowl**	0.547	0.178	10.163	**0.001**
			Terrestrial bird	-0.399	0.315	1.786	0.182
			Mammal	-0.007	0.089	0.005	0.941
0.1%	χ^2_{525} = 526.839	χ^2_{8} = 17.728	**Fish**	-0.241	0.130	3.781	0.052
	P = 0.469	*P* = 0.023	**Amphibian**	0.360	0.130	8.471	**0.004**
		R^2 = 7.03%	**Waterfowl**	0.544	0.180	9.813	**0.002**
			Terrestrial bird	-0.356	0.315	1.401	0.237
			Mammal	-0.036	0.092	0.157	0.692
0.5%	χ^2_{525} = 539.371	χ^2_{8} = 9.141	**Fish**	-0.331	0.155	5.150	**0.023**
	P = 0.323	*P* = 0.331	**Amphibian**	0.328	0.132	6.177	**0.013**
		R^2 = 9.91%	**Waterfowl**	0.633	0.180	12.400	**<0.001**
			Terrestrial bird	-0.962	0.465	5.714	**0.017**
			Mammal	0.067	0.108	0.380	0.538
1%	χ^2_{525} = 515.411	χ^2_{8} = 15.946	**Fish**	-0.547	0.206	9.077	**0.003**
	P = 0.609	*P* = 0.043	**Amphibian**	0.405	0.153	8.260	**0.004**
		R^2 = 14.45%	**Waterfowl**	0.654	0.210	11.246	**0.001**
			Terrestrial bird	-1.639	0.736	9.060	**0.003**
			Mammal	0.047	0.130	0.133	0.716
5%	Model could not be fit to the data.						
10%	χ^2_{525} = 0.405	χ^2_{8} = 0.382	**Fish**	-0.023	52.42	0.398	0.528
	P = 1.000	*P* = 1.000	**Amphibian**	0.039	11.63	162.241	**<0.001**
		R^2 = 98.83%	**Waterfowl**	0.091	15.65	0.920	0.338
			Terrestrial bird	3.971×10^3	2.536×10^7	3.559	0.059
			Mammal	-0.049	19.67	7.150	**0.008**
30%	Model could not be fit to the data.						
Species-specific	χ^2_{525} = 517.497	χ^2_{8} = 22.581	**Fish**	-0.238	0.124	4.049	**0.044**
	P = 0.584	*P* = 0.004	**Amphibian**	0.557	0.149	16.564	**<0.001**
		R^2 = 9.41%	**Waterfowl**	0.621	0.181	13.229	**<0.001**
			Terrestrial bird	-0.328	0.291	1.383	0.240
			Mammal	0.016	0.090	0.032	0.858

Table S5 Summary of different significant associations between great crested newt and other vertebrate species as determined by the probabilistic co-occurrence model at different metabarcoding sequence thresholds (*N* = 532 ponds).

Threshold	**Positive pairs**	**Negative pairs**	**Random pairs**	**Positive associations with great crested newt**		**Negative associations with great crested newt**	
				Species	***P***	**Species**	***P***
None	64	4	338	Cow	<0.001	Common carp	0.029
				Eurasian coot	0.007		
				Common moorhen	<0.001		
				Smooth newt	<0.001		
				Pig	<0.001		
0.05%	53	6	296	Eurasian coot	0.027	Toad	0.003
				Common moorhen	<0.001	Three-spined stickleback	0.003
				Smooth newt	<0.001	Grey squirrel	0.032
				Pig	0.002		
0.1%	47	7	277	Eurasian coot	0.032	Toad	0.011
				Common moorhen	0.001	Three-spined stickleback	0.009
				Smooth newt	<0.001	Grey squirrel	0.023
				Pig	0.009		
0.5%	37	13	205	Eurasian coot	0.008	Toad	0.006
				Common moorhen	0.001	Three-spined stickleback	0.009
				Smooth newt	<0.001	Grey squirrel	0.005
				Pig	0.004	Pike	0.031
						Common pheasant	0.023
1%	23	9	169	Common moorhen	0.001	Toad	0.010
				Smooth newt	<0.001	Three-spined stickleback	0.001
				Pig	0.014	Grey squirrel	0.042
						Pike	0.044
						Common pheasant	0.012
5%	3	7	76	Common moorhen	0.007	Toad	0.004
				Smooth newt	<0.001	Three-spined stickleback	0.004
						Common carp	0.029
10%	2	3	51	Smooth newt	<0.001	Toad	0.020
						Three-spined stickleback	0.003
30%	0	1	11				
Species-specific	48	17	299	Eurasian coot	0.023	Toad	0.009
				Common moorhen	0.001	Three-spined stickleback	0.009
				Smooth newt	< 0.001	Grey squirrel	0.018
				Pig	0.004	Common pheasant	0.048
						Ninespine stickleback	0.047

Table S6 Summary of abiotic and biotic determinants of great crested newt occupancy as identified using a binomial GLMM for different metabarcoding sequence thresholds (*N* = 504 ponds). For categorical variables with more than one level, effect size and standard error are only given for levels reported in the model summary. Test statistic is for LRT used. Significant P-values (<0.05) are in bold.

Threshold	**Overdispersion**	**Model fit**	**Model variables**	**Effect size**	**Standard error**	χ^2	***P***
No threshold	χ^2_{496} = 525.999	χ^2_{8} = 14.167	**Smooth newt**	1.303	0.252	29.174	**<0.001**
	P = 0.170	*P* = 0.078	**Species richness**	0.305	0.053	37.618	**<0.001**
		R^2 = 33.94%	**Inflow**	-0.757	0.244	10.029	**0.002**
			Ruderals			6.690	**0.035**
			None	-0.813	0.455		
			Some	-0.313	0.466		
			Common carp	-1.584	0.501	12.374	**<0.001**
0.05%	χ^2_{490} = 405.328	χ^2_{8} = 6.171	**Smooth newt**	0.635	0.278	5.794	**0.016**
	P = 0.998	*P* = 0.628	**Species richness**	0.510	0.104	52.263	**<0.001**
		R^2 = 40.99%	**Common toad**	-1.936	0.505	24.704	**<0.001**
			Grey squirrel	-2.140	0.603	19.946	**<0.001**
			Three-spined	-1.703	0.503	17.317	**<0.001**
			stickleback	-0.913	0.306	10.671	**0.001**
			Inflow	-	0.0002	5.726	**0.017**
			Pond area	0.0004	0.492	7.934	**0.047**
			Permanence	0.482			
			Never dries		0.539		
			Rarely dries	0.213	0.530		
			Sometimes dries	-0.420		6.055	**0.048**
			Ruderals		0.552		
			None	-0.567	0.551		
			Some	0.067			
0.1%	χ^2_{488} = 407.611	χ^2_{8} = 6.232	**Species richness**	0.510	0.115	82.906	**< 0.001**
	P = 0.997	*P* = 0.621	**Common toad**	-1.844	0.518	21.710	**<0.001**
		R^2 = 41.00%	**Inflow**	-0.866	0.311	9.350	**0.002**
			Grey squirrel	-2.386	0.666	20.517	**<0.001**
			Max. depth	0.403	0.143	9.144	**0.003**
			Three-spined	-1.623	0.495	16.589	**<0.001**
			stickleback	0.010	0.005	4.493	**0.034**
			Macrophytes	-	0.0002	7.730	**0.005**
			Pond area	0.0005		9.752	**0.008**
			Ruderals		0.542		
			None	-0.698	0.543		
			Some	0.107		7.375	**0.025**
			Woodland		0.366		
			None	-0.874	0.322		
			Some	-0.279		7.324	**0.026**
			Terrestrial other		0.456		
			None	0.322	0.446		
			Some	-0.402			

0.5%	χ^2_{491} = 352.876	χ^2_8 = 17.172	**Species richness**	0.739	0.158	83.028	**<0.001**
	P = 0.999	*P* = 0.028	**Common toad**	-2.227	0.641	23.505	**<0.001**
		R^2 = 47.27%	**Inflow**	-1.421	0.402	21.583	**<0.001**
			Pond area	-	0.0003	6.955	**0.008**
			Three-spined	0.0006	0.588	15.679	**<0.001**
			stickleback	-1.847		18.733	**<0.001**
			Permanence		0.543		
			Never dries	0.950	0.576		
			Rarely dries	0.689	0.574		
			Sometimes dries	-0.595	0.881	26.827	**<0.001**
			Grey squirrel	-3.126		9.606	**0.008**
			Woodland		0.401		
			None	-0.961	0.340		
			Some	-0.143			
1%	χ^2_{496} = 485.663	χ^2_8 = 5.940	**Species richness**	0.608	0.130	56.081	**<0.001**
	P = 0.622	*P* = 0.654	**Overhang**	-0.011	0.004	8.463	**0.004**
		R^2 = 38.34%	**Three-spined**	-2.132	0.632	20.225	**<0.001**
			stickleback	-	0.0002	10.201	**0.001**
			Pond area	0.0006	0.340	16.056	**<0.001**
			Inflow	-1.144	0.134	4.319	**0.038**
			Max. depth	0.266			
5%	Model could not be fit to the data.						
10%	No explanatory variables retained by model selection - null model had better fit than final model from model selection. Due to threshold stringency and highly reduced detection of great crested newt, no explanatory variables adequately fit the data.						
30%	No explanatory variables retained by model selection - null model had better fit than final model from model selection. Due to threshold stringency and highly reduced detection of great crested newt, no explanatory variables adequately fit the data.						
Species-specific	χ^2_{496} = 485.663	χ^2_8 = 5.940	**Smooth newt**	1.081	0.303	17.434	**<0.001**
	P = 0.622	*P* = 0.6540	**Species richness**	0.527	0.105	60.267	**<0.001**
		R^2 = 38.34%	**Common toad**	-1.635	0.696	8.228	**0.004**
			Grey squirrel	-1.591	0.534	12.432	**<0.001**
			Three-spined	-1.432	0.561	9.453	**0.002**
			stickleback	-	0.0002	6.453	**0.011**
			Inflow	0.0004	0.139	4.266	**0.039**
			Pond area	0.282			
			Pond depth		0.359	4.467	**0.035**
			Outflow	-0.713		6.507	6.507
			Ruderals		0.527		
			None	-0.617	0.528		
			Some	0.032		7.918	**0.019**
			Terrestrial other		0.429		
			None	0.428	0.424		
			Some	-0.316			

Table S7 Summary of relationship between HSI score and great crested newt occupancy as determined using a binomial GLMM for different metabarcoding sequence thresholds (*N* = 504 ponds). Test statistic is for LRT used. Significant P-values (<0.05) are in bold.

Threshold	GLMM results	Overdispersion	Model fit
None	2.649 ± 0.735 χ^2_1 = 13.791 ***P* < 0.001**	χ^2_{501} = 506.140 *P* = 0.428	χ^2_8 = 4.801 *P* = 0.779 R^2 = 3.88%
0.05%	3.070 ± 0.795 χ^2_1 = 16.114 ***P* < 0.001**	χ^2_{501} = 507.131 *P* = 0.415	χ^2_8 = 8.880 *P* = 0.353 R^2 = 5.14%
0.1%	3.081 ± 0.805 χ^2_1 = 15.831 ***P* < 0.001**	χ^2_{501} = 507.366 *P* = 0.412	χ^2_8 = 9.902 *P* = 0.272 R^2 = 5.18%
0.5%	3.3863 ± 0.841 χ^2_1 = 17.739 ***P* < 0.001**	χ^2_{501} = 510.637 *P* = 0.373	χ^2_8 = 14.558 *P* = 0.068 R^2 = 6.19%
1%	3.775 ± 0.887 χ^2_1 = 20.163 ***P* < 0.001**	χ^2_{501} = 511.628 *P* = 0.362	χ^2_8 = 16.657 ***P* = 0.034** R^2 = 7.58%
5%	Null model better fit to data. Great crested newt occupancy no longer explained by HSI score.		
10%	Null model better fit to data. Great crested newt occupancy no longer explained by HSI score.		
30%	Null model better fit to data. Great crested newt occupancy no longer explained by HSI score.		
Species-specific	3.020 ± 0.791 χ^2_1 = 15.709 ***P* < 0.001**	χ^2_{501} = 506.763 *P* = 0.420	χ^2_8 = 8.118 *P* = 0.422 R^2 = 4.99%

Table S8 Summary of abiotic and biotic determinants of vertebrate species richness as identified using a Poisson GLMM for different metabarcoding sequence thresholds (*N* = 504 ponds). For categorical variables with more than one level, effect size and standard error are only given for levels reported in the model summary. Test statistic is for LRT used. Significant P-values (<0.05) are in bold.

Threshold	**Model overdispersion**	**Model fit**	**Model variables**	**Effect size**	**Standard error**	χ^2	***P***
No threshold	χ^2_{498} = 375.433	χ^2_{8} = -69.777	**Overhang**	-0.002	0.001	10.935	**0.001**
	P = 0.999	*P* = 1.000	**Rough grass**			8.205	**0.017**
		R^2 = 6.66%	None	0.062	0.002		
			Some	-0.112	0.002		
			Outflow	0.200	0.002	10.988	**0.001**
0.05%	χ^2_{496} = 406.722	χ^2_{8} = -62.768	**Overhang**	-0.002	0.001	6.963	**0.008**
	P = 0.999	*P* = 1.000	**Outflow**	0.163	0.062	6.735	**0.010**
		R^2 = 6.68%	**Rough grass**			7.374	**0.025**
			None	0.009	0.068		
			Some	-0.145	0.065		
			Scrub/hedge			6.722	**0.035**
			None	-0.079	0.131		
			Some	0.139	0.057		
0.1%	χ^2_{496} = 410.479	χ^2_{8} = -62.194	**Overhang**	-0.002	0.001	8.628	**0.003**
	P = 0.998	*P* = 1.000	**Outflow**	0.161	0.063	6.443	**0.011**
		R^2 = 6.94%	**Rough grass**			6.538	**0.038**
			None	0.006	0.069		
			Some	-0.140	0.066		
			Scrub/hedge			6.891	**0.032**
			None	-0.091	0.134		
			Some	0.141	0.058		
0.5%	χ^2_{496} = 508.449	χ^2_{8} = -1.413	**Overhang**	-0.002	0.001	9.090	**0.003**
	P = 0.340	*P* = 1.000	**Outflow**	0.152	0.062	5.946	**0.015**
		R^2 = 6.54%	**Rough grass**			7.430	**0.024**
			None	-0.064	0.076		
			Some	-0.184	0.072		
			Overall terrestrial habitat			6.485	**0.039**
			Moderate	0.193	0.078		
			Poor	0.177	0.087		
1%	χ^2_{501} = 470.396	χ^2_{8} = -35.854	**Overhang**	-0.003	0.001	14.810	**<0.001**
	P = 0.833	*P* = 1.000					
		R^2 = 3.50%					
5%	χ^2_{499} = 378.448	χ^2_{8} = 39.565	**Overhang**	-0.004	0.001	16.921	**<0.001**
	P = 0.999	*P* = **<0.001**	**Rough grass**			8.126	**0.017**
		R^2 = 7.66%	None	0.061	0.092		
			Some	-0.185	0.093		

10%	$\chi^2_{501} = 357.332$	$\chi^2_8 = -238.540$	**Overhang**	-0.007	0.001	26.768	**<0.001**
	$P = 0.999$	$P = 1.000$					
		$R^2 = 7.68\%$					
30%	$\chi^2_{497} = 341.011$	$\chi^2_8 = 10.709$	**Overhang**	-0.011	0.002	25.478	**<0.001**
	$P = 1.000$	$P = 0.219$	**Waterfowl**			7.493	**0.024**
		$R^2 = 12.65\%$	Major	-1.169	0.513		
			Minor	-0.122	0.149		
			Woodland			6.289	**0.043**
			None	-0.448	0.185		
			Some	-0.146	0.179		
Species-	$\chi^2_{494} = 431.959$	$\chi^2_8 = -42.708$	**Outflow**	0.214	0.063	11.220	**0.001**
specific	$P = 0.979$	$P = 1.000$	**Rough grass**			16.715	**<0.001**
		$R^2 = 8.94\%$	None	-0.1402	0.0795		
			Some	-0.297	0.074		
			Overall terrestrial habitat			8.244	**0.016**
			Poor	0.115	0.089		
			Moderate	0.216	0.078		
			Overhang	-0.0026	0.0008	9.575	**0.002**
			Macrophyte cover	-0.002	0.001	4.117	**0.043**
			Pond density	0.006	0.003	4.564	**0.033**

Table S9 Summary of relationship between HSI score and vertebrate species richness as determined using a binomial GLMM for different metabarcoding sequence thresholds (*N* = 504 ponds). Test statistic is for LRT used. Significant P-values (<0.05) are in bold.

Threshold	**GLMM results**	**Overdispersion**	**Model fit**
None	0.474 ± 0.192 χ^2_1 = 6.102 ***P* = 0.014**	χ^2_{501} = 355.432 *P* = 0.999	χ^2_8 = -109.49 *P* = 1.000 R^2 = 1.29%
0.05%	0.496 ± 0.002 χ^2_1 = 6.244 ***P* = 0.013**	χ^2_{501} = 380.354 *P* = 0.999	χ^2_8 = -125.06 *P* = 1.000 R^2 = 1.35%
0.1%	0.504 ± 0.002 χ^2_1 = 6.251 ***P* = 0.012**	χ^2_{501} = 382.557 *P* = 0.999	χ^2_8 = -130.31 *P* = 1.000 R^2 = 1.36%
0.5%	0.472 ± 0.198 χ^2_1 = 5.732 ***P* = 0.017**	χ^2_{501} = 447.442 *P* = 0.769	χ^2_8 = -42.281 *P* = 1.000 R^2 = 1.32%
1%	0.561 ± 0.210 χ^2_1 = 7.267 ***P* = 0.007**	χ^2_{501} = 473.185 *P* = 0.809	χ^2_8 = -5.908 *P* = 1.000 R^2 = 1.73%
5%	0.683 ± 0.277 χ^2_1 = 6.193 ***P* = 0.013**	χ^2_{501} = 389.934 *P* = 0.999	χ^2_8 = -47.496 *P* = 1.000 R^2 = 1.64%
10%	0.897 ± 0.336 χ^2_1 = 7.292 ***P* = 0.007**	χ^2_{501} = 370.163 *P* = 0.999	χ^2_8 = 126.330 ***P* < 0.001** R^2 = 2.13%
30%	1.189 ± 0.546 χ^2_1 = 4.894 ***P* = 0.027**	χ^2_{501} = 350.580 *P* = 0.999	χ^2_8 = 10.472 *P* = 0.233 R^2 = 2.03%
Species-specific	0.459 ± 0.002 χ^2_1 = 4.894 ***P* = 0.025**	χ^2_{501} = 389.744 *P* = 0.999	χ^2_8 = -145.120 *P* = 1.000 R^2 = 1.10%

Table S10 Summary of species detected by eDNA metabarcoding of freshwater ponds (N = 532).

Common name	**Binomial name**	**No. ponds detected**
European eel	*Anguilla anguilla*	15
Common barbel	*Barbus barbus*	2
Crucian carp	*Carassius carassius*	2
Common carp	*Cyprinus carpio*	41
Common minnow	*Phoxinus phoxinus*	13
Common roach	*Rutilus rutilus*	72
European chub	*Squalius cephalus*	21
Stone loach	*Barbatula barbatula*	15
Northern pike	*Esox lucius*	17
European bullhead	*Cottus gobio*	14
Three-spined stickleback	*Gasterosteus aculeatus*	56
Ninespine stickleback	*Pungitius pungitius*	15
Ruffe	*Gymnocephalus cernua*	1
Rainbow trout	*Oncorhynchus mykiss*	3
Common toad	*Bufo bufo*	42
Marsh frog	*Pelophylax ridibundus*	1
Common frog	*Rana temporaria*	120
Palmate newt	*Lissotrition helveticus*	5
Smooth newt	*Lissotriton vulgaris*	152
Great crested newt	*Triturus cristatus*	149
Green-winged teal	*Anas carolinensis*	7
Eurasian oystercatcher	*Haematopus ostralegus*	1
Common buzzard	*Buteo buteo*	4
Common pheasant	*Phasianus colchicus*	25
Domesticated turkey	*Meleagris gallopavo*	11
Helmeted guineafowl	*Numida meleagris*	1
Eurasian coot	*Fulica atra*	48
Common moorhen	*Gallinula chloropus*	215
Eurasian jay	*Garrulus glandarius*	7
European goldfinch	*Carduelis carduelis*	1

Dunnock	*Prunella modularis*	4
Eurasian nuthatch	*Sitta europaea*	1
Common starling	*Sturnus vulgaris*	4
Melodius warbler	*Hippolais polyglotta*	2
Grey heron	*Ardea cinerea*	1
Great spotted woodpecker	*Dendrocopus major*	1
Green woodpecker	*Picus viridis*	2
Tawny owl	*Strix aluco*	1
Dog	*Canis lupus*	65
Red fox	*Vulpes vulpes*	9
Eurasian otter	*Lutra lutra*	1
European badger	*Meles meles*	7
European polecat	*Mustela putorius*	1
Common pipistrelle	*Pipistrellus pipistrellus*	1
Eurasian water shrew	*Neomys fodiens*	9
Common shrew	*Sorex araneus*	1
European hare	*Lepus europaeus*	1
European rabbit	*Oryctolagus cuniculus*	24
Horse	*Equus caballus*	3
European water vole	*Arvicola amphibius*	16
Bank vole	*Myodes glareolus*	9
House mouse	*Mus musculus*	16
Brown rat	*Rattus norvegicus*	39
Grey squirrel	*Sciurus carolinensis*	57
Cow	*Bos taurus*	179
Sheep	*Ovis aries*	42
Red deer	*Cervus elaphus*	2
Reeve's muntjac	*Muntiacus reevesi*	3
Pig	*Sus scrofa*	140
Cat	*Felis catus*	16

Appendix 4: Figures

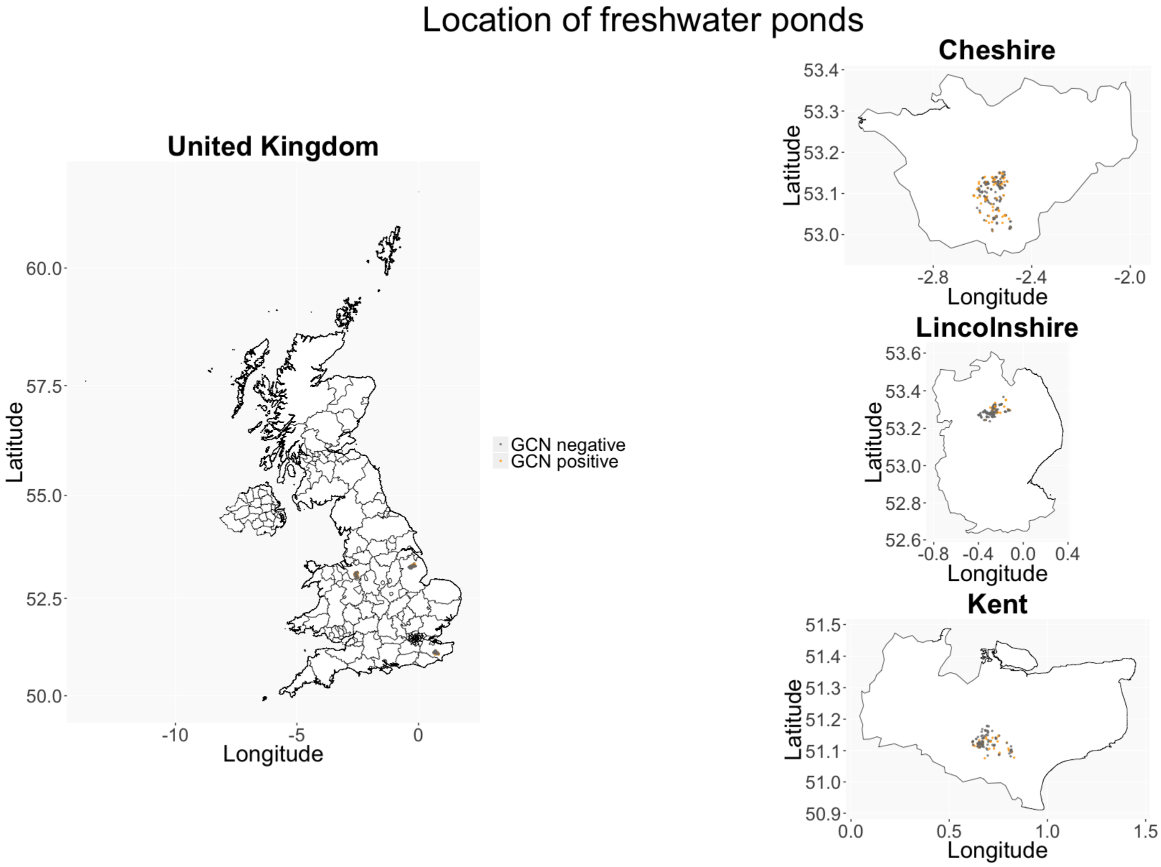

Figure S1 Location of ponds (*N* = 504) sampled for eDNA as part of Natural England's Great Crested Newt Evidence Enhancement Programme. Ponds that were negative or positive for great crested newt (GCN) by targeted qPCR are indicated by grey and orange points respectively.

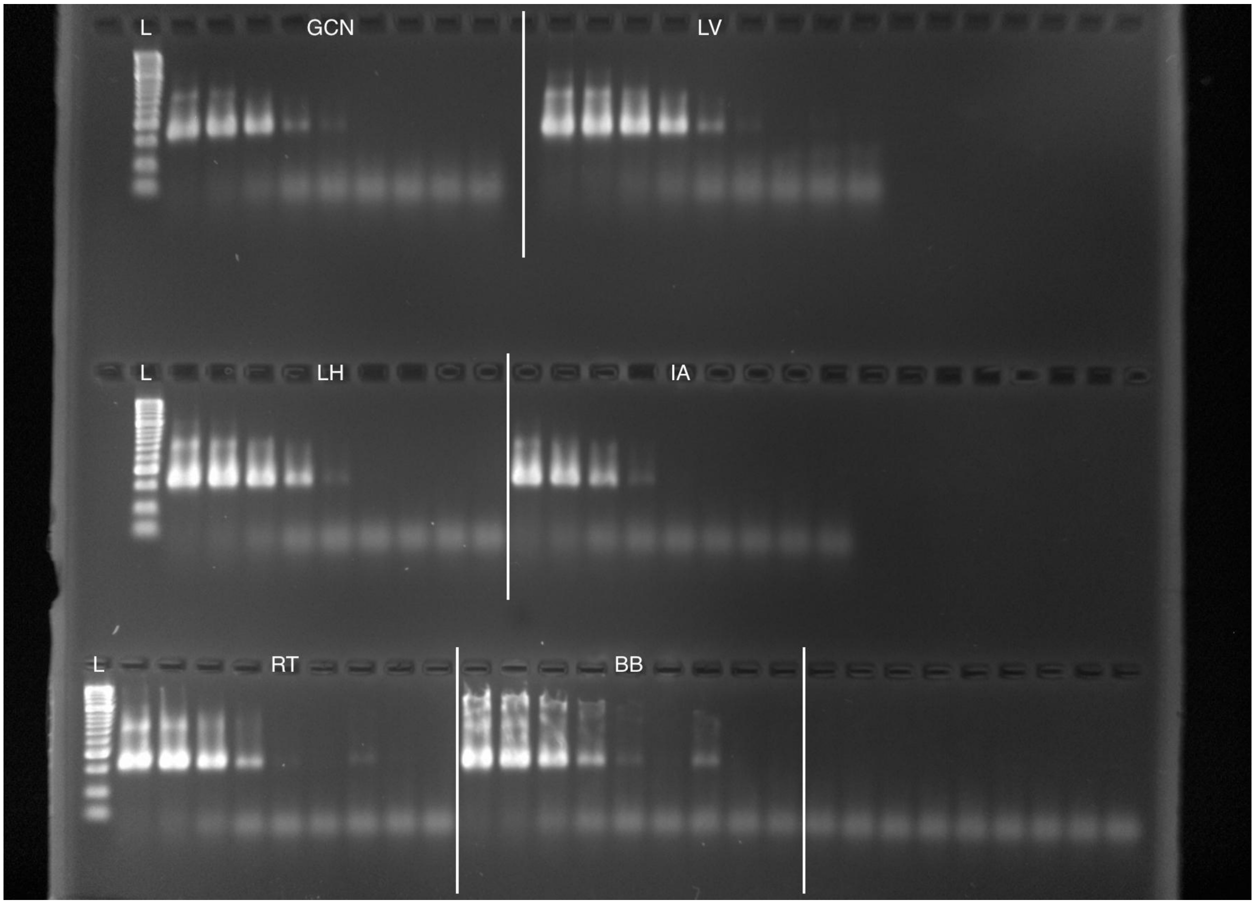

Figure S2 Gel image showing results of *in vitro* primer validation. All tissue DNA used for dilution series was standardised to a starting concentration of 5 ng/µl. The LOD was variable for each species: great crested newt (GCN), palmate newt (LH), common frog (RT) and common toad (BB) were not amplified below 5 x 10^{-4} ng/µl, whereas Alpine newt (IA) was was not amplified below 5 x 10^{-3} ng/µl and smooth newt (LV) below 5 x 10^{-5} ng/µl.

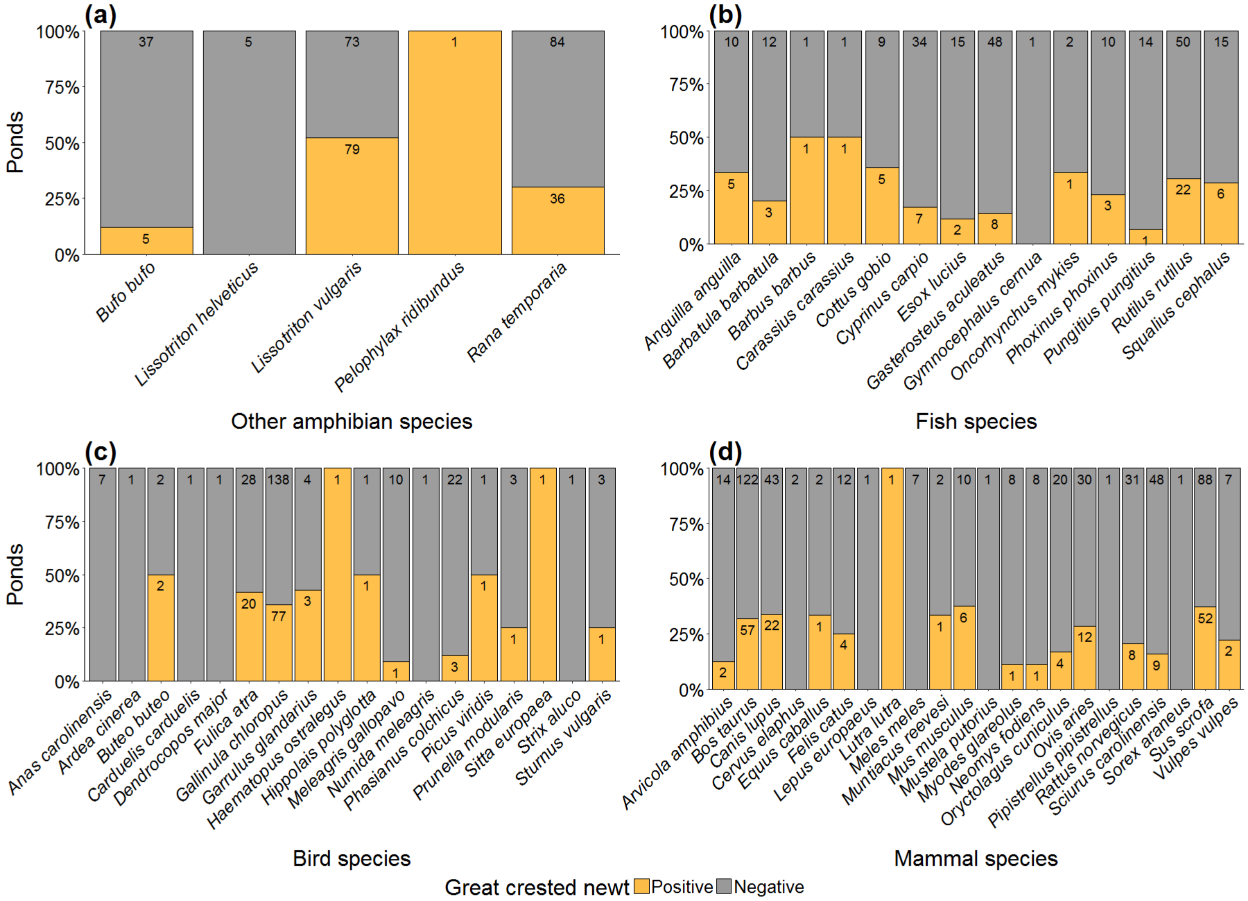

Figure S3 Occurrence of great crested newt in relation to species from different vertebrate groups (*N* = 532 ponds): (a) other amphibians, (b) fish, (c) birds, and (d) mammals. Numbers on each bar are the number of ponds in which a species was detected with and without great crested newt respectively.

References

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R. A., … Dunn, F. (2015). Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (*Triturus cristatus*). *Biological Conservation*, *183*, 19–28. https://doi.org/10.1016/j.biocon.2014.11.029

Bjørnstad, O. N. (2009). ncf: spatial nonparametric covariance functions. R package version 1.1-7.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Edgar, P., & Bird, D. R. (2006). *Action plan for the conservation of the crested newt Triturus cristatus species complex in Europe*. Council of the European Union, Strassbourg, Germany.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*, 2194–2200. https://doi.org/10.1093/bioinformatics/btr381

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., … Pompanon, F. (2010). An *In silico* approach for the evaluation of DNA barcodes. *BMC Genomics*, *11*, 434. https://doi.org/10.1186/1471-2164-11-434

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression, Second Edition*. Sage, Thousand Oaks, CA.

Griffith, D., Veech, J., & Marsh, C. (2016). cooccur: Probabilistic Species Co-Occurrence Analysis in R. *Journal of Statistical Software*, *69*, 1–17. https://doi.org/10.18637/jss.v069.c02

Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., … Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, *25*, 3101–3119. https://doi.org/10.1111/mec.13660

Harper, L. R., Lawson Handley, L., Hahn, C., Boonham, N., Rees, H. C., Gough, K. C., … Hänfling, B. (2018). Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (*Triturus cristatus*). *Ecology and Evolution*, *8*, 6330–6341. https://doi.org/10.1002/ece3.4013

Haubrock, P. J., & Altrichter, J. (2016). Northern crested newt (*Triturus cristatus*) migration in a nature reserve: multiple incidents of breeding season displacements exceeding 1km. *The Herpetological Bulletin*, *138*, 31–33.

Kitson, J. J. N., Hahn, C., Sands, R. J., Straw, N. A., Evans, D. M., & Lunt, D.H. (2018). Detecting host-parasitoid interactions in an invasive Lepidopteran using nested tagging DNA-metabarcoding. *Molecular Ecology.* https://doi.org/10.1111/mec.14518

Kozlov, A. M., Zhang, J., Yilmaz, P., Glöckner, F. O., & Stamatakis, A. (2016). Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, *44*, 5022–5033. https://doi.org/10.1093/nar/gkw396

Larsson, A. (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*, 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Magoč, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*, 2957–2963. https://doi.org/10.1093/bioinformatics/btr507

O'Donnell, J. L., Kelly, R. P., Lowell, N. C., & Port, J. A. (2016). Indexed PCR Primers Induce Template-Specific Bias in Large-Scale DNA Sequencing Studies. *PLoS ONE*, *11*, e0148698. https://doi.org/10.1371/journal.pone.0148698

Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., … Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, *25*, 527–541. https://doi.org/10.1111/mec.13481

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 Research*, *39*, e145. https://doi.org/10.1093/nar/gkr732

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*, 2688–2690. https://doi.org/10.1093/bioinformatics/btl446

Szitenberg, A., John, M., Blaxter, M. L., & Lunt, D. H. (2015). ReproPhylo: An Environment for Reproducible Phylogenomics. *PLoS Computational Biology*, *11*, e1004447. https://doi.org/10.1371/journal.pcbi.1004447

Therneau, T., Atkinson, B. & Ripley, B. (2014). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13.

Thomsen, P. F., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., Orlando, L., & Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, *21*, 2565–2573. https://doi.org/10.1111/j.1365-294X.2011.05418.x

Veech, J.A. (2013). A probabilistic model for analysing species co-occurrence. *Global Ecology and Biogeography*, *22*, 252–260. https://doi.org/10.1111/j.1466-8238.2012.00789.x

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, *7*, 203–214. https://doi.org/10.1089/10665270050081478

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer New York, USA.