Received: 24 May 2020 | Revised: 7 January 2021 | Accepted: 24 February 2021

DOI: 10.1111/1755-0998.13367

INVITED TECHNICAL REVIEW

MOLECULAR ECOLOGY RESOURCES WILEY

Improving the reliability of eDNA data interpretation

Alfred Burian[1,2,3] | Quentin Mauvisseau[1,4] | Mark Bulling[1] | Sami Domisch[5] | Song Qian[6] | Michael Sweet[1]

[1]Aquatic Research Facility, Environmental Sustainability Research Centre, University of Derby, Derby, UK

[2]Marine Ecology Department, Lurio University, Nampula, Mozambique

[3]Department of Computational Landscape Ecology, UFZ – Helmholtz Centre for Environmental Research, Leipzig, Germany

[4]Natural History Museum, University of Oslo, Oslo, Norway

[5]Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

[6]Department of Environmental Sciences, University of Toledo, Toledo, OH, USA

Correspondence
Quentin Mauvisseau, Aquatic Research Facility, Environmental Sustainability Research Centre, University of Derby, Derby, UK.
Email: quentin.mauvisseau@nhm.uio.no

Funding information
Global Challenges Research Fund, UK; Leibniz Competition, Grant/Award Number: J45/2018

Abstract

Global declines in biodiversity highlight the need to effectively monitor the density and distribution of threatened species. In recent years, molecular survey methods detecting DNA released by target-species into their environment (eDNA) have been rapidly on the rise. Despite providing new, cost-effective tools for conservation, eDNA-based methods are prone to errors. Best field and laboratory practices can mitigate some, but the risks of errors cannot be eliminated and need to be accounted for. Here, we synthesize recent advances in data processing tools that increase the reliability of interpretations drawn from eDNA data. We review advances in occupancy models to consider spatial data-structures and simultaneously assess rates of false positive and negative results. Further, we introduce process-based models and the integration of metabarcoding data as complementing approaches to increase the reliability of target-species assessments. These tools will be most effective when capitalizing on multi-source data sets collating eDNA with classical survey and citizen-science approaches, paving the way for more robust decision-making processes in conservation planning.

KEYWORDS
barcoding, Bayesian analysis, data fusion, detection probability, eDNA, false positives, metabarcoding, occupancy modelling, sources of error, species distribution modelling

1 | INTRODUCTION

Since the beginning of the last century, global species extinctions have occurred at unprecedented rates and currently over a million species are at risk (IPBES, 2019). Rapid transformations of natural ecosystems and habitat degradation highlight the urgent need for effective conservation strategies to mitigate further biodiversity losses. A prerequisite for the development of such strategies is the provision of comprehensive, reliable and frequently updated monitoring data, recording distribution changes of vulnerable, endangered and invasive species.

A promising approach that is currently gaining momentum and fulfils such monitoring objectives relies on the detection of genetic traces left by organisms, also referred to as environmental DNA (or eDNA). Substantial advantages of eDNA-based methods are higher cost and time effectiveness compared to many traditional survey methods (Evans et al., 2017), their noninvasive nature (Cristescu & Hebert, 2018), and high specificity and sensitivity (Wilcox et al., 2013). However, eDNA-based methods have only been applied for about a decade in conservation management, and method reliability and accuracy are still being refined (Cristescu & Hebert, 2018).

Alfred Burian and Quentin Mauvisseau contributed equally.

2 | SOURCES OF ERROR AND ERROR TYPES

Although there are many studies demonstrating the potential of eDNA-based methods, eDNA surveys are, like any sampling approach, prone to the emergence of errors (Doi et al., 2019; Ficetola et al., 2015). In general, two error types can be distinguished: false negatives (i.e., failures to detect a species despite its presence) and false positives, which are defined as the false detection of a species despite its absence in the field (Guillera-Arroita et al., 2017; Hansen et al., 2018; Lahoz-Monfort et al., 2016). Recent studies have suggested strategies to account for methodological (Zinger et al., 2019) and systematic sources of error (Miller et al., 2011). However, in many cases, only a subset of potential confounding factors are simultaneously considered. In this paper, we re-emphasize early arguments (Darling & Mahon, 2011) that differentiate among error types (false positives vs. false negatives), and the process level at which they occur (i.e., field or laboratory sample processing) is a critical step in the implementation of effective error mitigation protocols. Nevertheless, we clearly highlight that even best field and laboratory practices cannot remove all sources of error. Therefore, in the second part of this article, we focus on the use of statistical tools such as occupancy and process-based models to mitigate errors in single-species targeted and multispecies metabarcoding approaches. Combined with rapid increases in data availability due to lower analytical costs, these tools are already showing promise in providing enough leverage to lift eDNA-based applications in terrestrial and aquatic environments to a new level.

3 | DIFFERENTIATION OF SOURCES OF ERROR AND THEIR MITIGATION

We will differentiate here between sources of errors emerging at the laboratory and the field processing levels and highlight their potential to trigger false negative or false positive results (Table 1). Such differentiation is equally valid for terrestrial and aquatic systems as well as single (targeted) and multispecies (metabarcoding) sequencing approaches. Multispecies approaches are, however, susceptible to additional bioinformatical challenges, which have been extensively covered elsewhere (e.g. Callahan et al., 2016; Zinger et al., 2019).

Most methodological eDNA studies that aim to improve field sampling methods focus on specific technical aspects of sampling protocols, for instance addressing flocculation vs. filtration methods to concentrate eDNA, filter pore sizes, sample preservation or

TABLE 1 Potential sources of error of eDNA-based methods which emerge at the field or laboratory process levels and culminate either in false positive or false negative result. Descritption of the error sources and their mitigation potential are displayed

			Error description	Mitigation potential
Incorrect result	Field	False negative	• **Inappropriate sampling method** (wrong habitats within sites or insufficient water quantities are sampled)	Very high
			• **Environmental heterogeneity** (of environment and populations)	intermediate
			• **DNA persistence** (high microbial activity leads to rapid eDNA degradation and low concentrations)	intermediate
			• **Physiological factors** (e.g. seasonality in activity patterns; spawning)	high
			• **Hydrological drivers** (e.g. rain events can flush out eDNA)	intermediate
		False positive	• **Contamination during sampling** (accidential eDNA across site transfer)	Very high
			• **Sampling independent contamination** (e.g. by birds, predator faeces, ballast water or other human activities)	intermediate
			• **Vertical transport** (e.g. donwnstream eDNA along river networks; influenced by hydrology	intermediate
			• **DNA persistence** (e.g. historical eDNA resuspended from sediments)	high
	Laboratory	False positive	• **Technical contamination** (contamination across samples due to inappropriate lab procedures or mislabelling of samples)	Very high
			• **Insufficient specificity** (non-target eDNA is amplified)	high
		False negative	• **Inhibition** (amplification of target DNA is inhibited by co-extracted compounds)	intermediate
			• **Inadequate extraction protocol** (low extracted DNA concentrations)	Very high
			• **Inadequate sample conservation** (if extraction is delayed)	Very high
			• **Low sensitivity** (low amplification efficiency, high limit of detection)	Very high
			• **Insufficient specificity** (assays fails to detect certain haplotypes)	high

water collation strategies (Deiner et al., 2018; Li et al., 2018; Spens et al., 2017). These are important contributions improving method reliability, but it is crucial to acknowledge that incorrect results can frequently emerge from mechanisms that are relatively independent of sampling and laboratory protocols, and which can only partly be mitigated (Table 1).

One important factor leading to false negatives is a highly stochastic distribution of eDNA in the field emerging from environmental and ecological drivers (Figure 1). For example, low eDNA concentrations of a target species and small-scale heterogeneity in its distribution can reduce the probability of collecting target eDNA in any given sample (Case 1 in Figure 1). A mitigation strategy to account for resulting false negatives is to increase the sampling effort and adjust the number of natural replicates (independent eDNA samples taken per site) and technical qPCR replicates (Mauvisseau et al., 2019). Further, environmental heterogeneity (Case 2 in Figure 1) can partly be accounted for by considering the small-scale habitat requirements of target species and increasing the number of sampling sites (Troth et al., 2021).

Further, the risk of false negatives is exacerbated by external environmental conditions (Table 1). For example, false negatives can be generated by weather events, such as rainfall or storms, effectively reducing eDNA turnover times (higher flow rates in aquatic environments, washing away of eDNA in terrestrial environments; Sales et al., 2020). Likewise, seasonal or species-specific physiological factors are known to affect species activity and eDNA shedding rates (Buxton et al., 2017; Wood et al., 2019). This will affect detection probabilities and may increase the risk of false negatives during parts of the target species' annual cycle (Troth et al., 2021). eDNA turnover is also strongly impacted by microbial activity and ultraviolet radiation (Buxton et al., 2017), both of which vary seasonally, and may result in decreased detection probabilities. However, low turnover rates can also trigger false positive results as slow eDNA degradation, or resuspension of historical eDNA can wrongly indicate the presence of populations that went extinct or emigrated from sampling sites (Goldberg et al., 2018).

Another factor which can generate false presence indications is downstream transportation of eDNA in rivers (Case 3 in Figure 1). Studies have demonstrated that eDNA can be detected at distances of greater than 10 km, and potentially up to 100 km, downstream from source populations (Pont et al., 2018). The degree of influence of downstream transportation is determined by a range of factors including eDNA turnover times, shedding rates and the size of source populations (e.g., highlighted in Buxton et al., 2017). Finally, false positive results can also result from sampling-independent introduction of target eDNA into unoccupied habitats. Such "contamination" can either be caused by human activities (e.g., release of ballast water from ships) or naturally via, for example, faeces of the target species' predators (Case 4, Figure 1; Merkes et al., 2014).

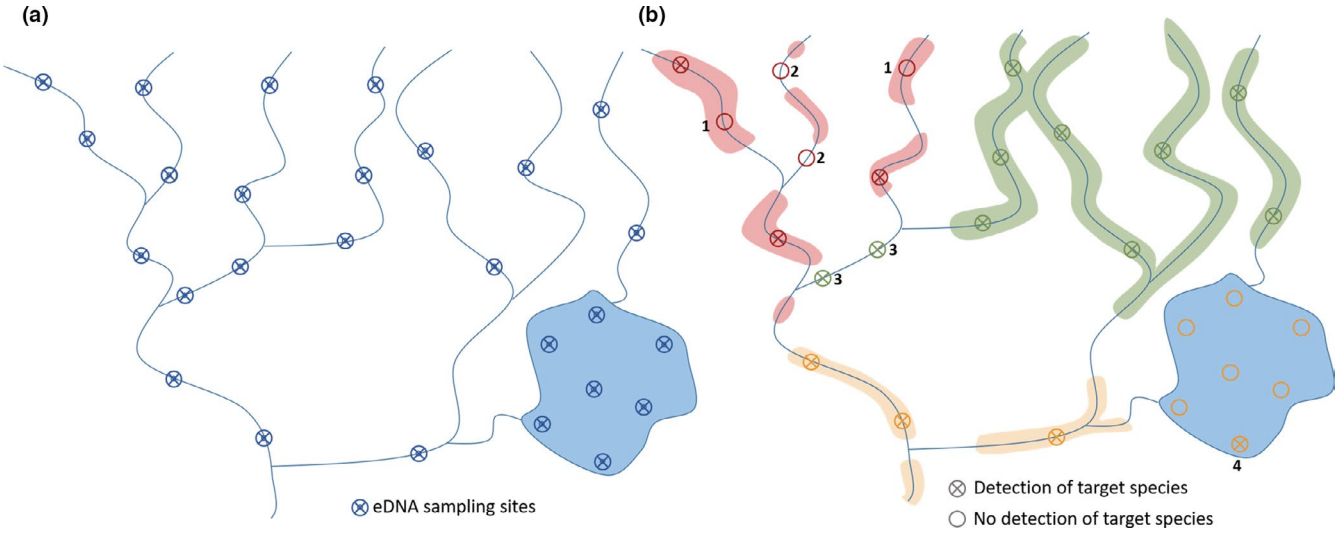


FIGURE 1 (a) A hypothetical river system sampled for eDNA to explore the presence/absence of three different 'target' species, an endemic and endangered rhithral species (green), an invasive rhithral species (red) and a native potamal species (yellow). (b) Illustration of true distributions of the three species (coloured river sections) and results of the eDNA sampling (empty circle and full circle represent negative and positive results, respectively). Numbers highlight four mismatches between actual species distributions and eDNA-based surveys. Case #1: A false negative occurs either due to low eDNA concentrations (e.g. high flow velocities and low shedding rates) or inhibition in the headwaters. Case #2: The endangered target species (red) is present, but survey results indicate its absence in the river section due to environmental heterogeneity and poor selection of sampling sites. Case #3: A river section is wrongly assessed as occupied by an invasive species. Here, eDNA from an upstream population is transported downstream leading to ecologically incorrect conclusions. Case #4: Horizontal eDNA transfer resulting from predator faeces or non-sampling-related human activities culminates in a false-positive detection. All four error examples are largely independent of sampling and laboratory protocols but can be mitigated with data processing tools. Cases 1, 2 and 4 can emerge in freshwater, marine and terrestrial environments. Case 3 is restricted to areas with directional eDNA transport (e.g. via wind in terrestrial or currents in marine systems)

In the course of laboratory analyses, false positives mainly result from (i) technical contamination through inappropriate procedures and (ii) nontarget eDNA triggering detection (i.e., insufficient specificity; Goldberg et al., 2016; Wilcox et al., 2013). Failures to detect eDNA, on the other hand, can emerge from a diverse range of sources including inadequate storage of samples or extraction protocols (Goldberg et al., 2016), DNA degradation after extraction, inhibition by co-extracted compounds, low sensitivity (i.e., failure to detect low eDNA concentrations; Klymus et al., 2019) and insufficient specificity (e.g., not all genetic variants of a species trigger positive results; Mauvisseau et al., 2019).

Recommendations to reduce the impact of these laboratory-based sources of error focus on the implementation of appropriate laboratory procedures and improvements of method sensitivity (Ficetola et al., 2016; Goldberg et al., 2016). Important measures include the optimization of primer design (Klymus et al., 2019) using negative controls at multiple levels (e.g., extraction of blank samples, and post-extraction controls; Ficetola et al., 2016) and thorough *in silico*, *in vitro* and *in vivo* testing (Ficetola et al., 2016)—based on established guidelines (e.g., Limit of Detection and Limit of Quantification; Klymus et al., 2019). Finally, inhibition of target eDNA is a frequently encountered challenge closely linked to the sampling environment (e.g., turbidity; Goldberg et al., 2016). Inhibition can be detected by adding synthetic DNA as an internal control (Klymus et al., 2019) and can be mitigated using inhibitor removal kits or by diluting DNA templates (Goldberg et al., 2016). However, both approaches also reduce concentrations of target eDNA and therefore inadvertently have the potential to increase the probability of false negative results.

This short synthesis of common sources of error clearly highlights that methodological optimizations and the reliance on sound ecological background knowledge are crucial for eDNA-based applications. However, a diverse set of errors can only partially be mitigated (Table 1) and will partly persist even when best practices are applied (Lahoz-Monfort et al., 2016). As already low error rates can severely affect the interpretation of results (Ruiz-Gutierrez et al., 2016), we want to next indicate how analytical tools can be utilized, complementary to those practices identified above, to increase the reliability of the use of eDNA-based methods.

4 | DATA PROCESSING TOOLS TO ACCOUNT FOR SOURCES OF ERRORS

Two powerful tools to account for emerging errors in survey data are the application of hierarchical occupancy and process-based models. These frameworks take different approaches, but both may estimate uncertainties related to eDNA-based species detection. They attempt to account for the probability of false negative and/or false positive results, thus increasing the information content of survey data and facilitating better-informed decision-making processes in ecosystem management and conservation.

4.1 | Occupancy models

Occupancy modelling has developed primarily from statistical approaches to model species distributions accounting for false negatives (Guillera-Arroita, 2017). They are based on a hierarchical structure, recognizing that the probability of detecting a species is contingent on the species being present (see Web Panel 1 for a brief and basic introduction). Models therefore evaluate occupancy probabilities (i.e., the probability that the target species is present at a given site) and detection probabilities (i.e., the probability of detection given that the species is present) as responses to environmental factors or/and co-occurrence of other species (Goldberg et al., 2018; Orzechowski et al., 2019). Specific eDNA occupancy models also evaluate a third probability, the probability of eDNA capture (Doi et al., 2019). Capture probabilities account for the chance of collecting target eDNA in a natural replicate, while detection probabilities denote the probability of detecting captured eDNA with one technical (PCR) replicate. This facilitates the consideration of complex ecological and environmental interactions, at least when sufficiently large data sets of presence–absence records are available to support model structures (Mackenzie & Royle, 2005). One fundamental data requirement is the availability of multiple observations per site (at least two) within a given time period of assumed constant occupancy (referred to as the "closure assumption"; Rota et al., 2009). eDNA-based assays incorporate the simultaneous collection of multiple natural replicates per site as a standard approach, and consequently occupancy models represent a very well-matched tool to increase the reliability and applicability of such data (Brost et al., 2018).

Occupancy models can either be based on frequentist or Bayesian statistical frameworks (Bailey et al., 2014; Ferguson et al., 2015). The main implementation difference between the two lies in the computational methods for parameter estimation: whilst frequentist approaches apply maximum likelihood estimation, Bayesian models are in most cases based on Markov chain Monte Carlo simulation (MCMC) procedures (Web Panel 1). Both frameworks can be implemented using various platforms (summarized in Table S1). A major advantage of Bayesian models is the possibility to include prior information (e.g., expert opinion on the likelihood of species presence) in the process of parameter estimation (Griffin et al., 2019). Further, Bayesian approaches are characterized by a higher inherent flexibility supporting models of greater real-world complexity (Guillera-Arroita, 2017), which has driven the recent surge in their application (Dorazio & Erickson, 2018; Orzechowski et al., 2019). However, they are computationally demanding and can, at times, differ in the data requirement needed to establish models of any given complexity. Consequently, the choice of framework and platform used in an eDNA context should be adjusted in accordance with data characteristics as well as management and scientific objectives.

Up to now, most occupancy models used in an eDNA context have been applied to single-species data, with multispecies approaches starting to be developed more recently (Doi et al., 2019; McClenaghan et al., 2020). Such multispecies metabarcoding-based

BOX 1 Capitalizing on metabarcoding community data to correct for false negatives

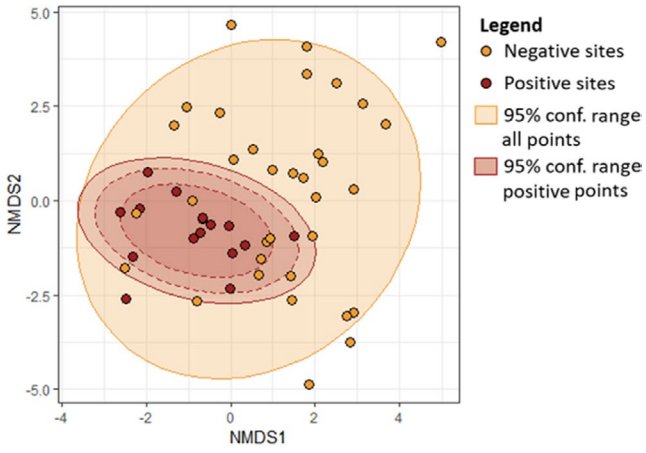
Conservation and ecosystem management requires reliable monitoring of overall biodiversity as well as species of particular interest. Metabarcoding approaches are powerful tools to capture species diversity but they are hampered by higher false negative rates compared to targeted approaches (Harper et al., 2018), resulting in methodological trade-offs .

Here we assume that a conservation manager needs to monitor biodiversity *and* the distribution of an endangered species and choose a metabarcoding approach to assess both. A potential tool to correct for false negative measurements of the target species is the use of community data to detect them. In an eDNA context, such tools are largely underexplored. Therefore, we introduce here two options to account for species interactions in occupancy models using metabarcoding data.

Option 1—PCA approach: Co-occurring species can affect the distribution (through, for example, competition or predation) and the detection (induced through behavioural changes or changes in relative abundance) of the target species. One possibility to integrate such effects into occupancy models is to condense the information recorded in the community matrix (ASV table) using a principal component analysis (PCA). Single PCs can then be included alongside environmental variables as predictors of the target species' occupancy and detection probability. The number of included PCs would then depend on the size of the data set and the variance they explain. This approach is simple to implement and suitable for small data sets. However, the establishment of PCs is untargeted, and the condensed information does not necessarily relate directly to the target species.

Option 2—targeted beta diversity approach: A more focused but currently still more experimental approach is the inclusion of a site-specific prior into Bayesian occupancy models. A site-specific prior score reflects how likely the target species is to occur at a site with a given community composition and is established in a two-step procedure. First, the overall strength of the prior (flat vs. strong) across all sites needs to be determined. This can be accomplished by calculating a community similarity matrix, which is condensed into a nonmetric multidimensional scaling (NMDS) plot (see below). If positive sites are tightly clustered, overall community composition is strongly associated with the target species' occurrence and the prior score needs to differ substantially across sites. The clustering of positive sites can be determined by dividing the 95% range of all positive sites (range that probably contains 95% of all positive sites; red solid line) by the 95% range of all sites (yellow line). The inverse of this fraction can then be used to determine the prior range in the occupancy model.

In a second step, an individual prior score for each site needs to be established. This can be achieved using a density probability function (e.g., based on the log-distance to the centroid of all positive sites) to determine the likelihood of each negative site being a false negative. This likelihood can be standardized by the prior range and then be included in the occupancy model. The utility of this approach still awaits testing, but it provides the substantial advantage of facilitating the incorporation of complex targeted information in occupancy model frameworks without substantially increasing data requirements.



assessments have the major advantage of providing information on biodiversity and community composition. However, their drawbacks include higher rates of false positive (Ficetola et al., 2016; Zinger et al., 2019) and false negative results (Harper et al., 2018). Thanks to bioinformatic or methodological advancements, false positives emerging from, for example, tag-jumps, formation of chimeric fragments and reagent contaminants can be mitigated (Schnell et al., 2015; Zinger et al., 2019; Zizka et al., 2019). False negative rates, on the other hand, represent a major challenge (naïve occupancy rates can be half that of targeted approaches; Harper et al., 2018), highlighting the necessity to account for them with data-processing tools.

4.2 | Computational "add-ons" to occupancy models

A potentially powerful option to mitigate higher false negative rates is the consideration of wider community data as co-existing species often shape the realized niche of the target organisms (Box 1). Recently developed multispecies occupancy models for eDNA data (Doi et al., 2019; McClenaghan et al., 2020) are an important advancement but do not account for species interactions in their model structure. The other end of the spectrum is represented by joint distribution models, which incorporate complex interactions among species or functional groups but which are often highly complex and data hungry (Pollock et al., 2014). A middle way that facilitates the consideration of species interactions, or at least the co-occurrence of their DNA at a sampling location without requiring hundreds of data points, is the use of community similarity indices, which we introduce in Box 1. Their performance could be further improved by including additional variables such as sequencing depth (McClenaghan et al., 2020) or absolute abundance measures (e.g., acoustic surveys of fish biomass) accounting for the dependency of false negative rates on total eDNA densities in occupancy models.

A major current challenge for occupancy models is that most do not account for false positives (Ferguson et al., 2015), although even low levels of false positives can substantially affect the reliability of model predictions (e.g., 2%–3% false positives may result in 50% overestimation of occurrence; Ruiz-Gutierrez et al., 2016). Further, the impact of non-accounted false positives is dependent on the number of natural replicates used in a survey (Ficetola et al., 2015). Normally, one would expect that the accuracy of model predictions will be positively affected by a higher number of natural replicates. However, once false positive rates increase, the gain provided by the higher number of natural replicates quickly disappears and is turned into a negative effect (Figure 2). These nontrivial relationships can be partly compensated for by setting detection thresholds (Ficetola et al., 2015) but, if overlooked, they can result in major flaws in sampling design and strategies.

A key difficulty in developing occupancy models correcting for false positives (emerging from either a method or a process type

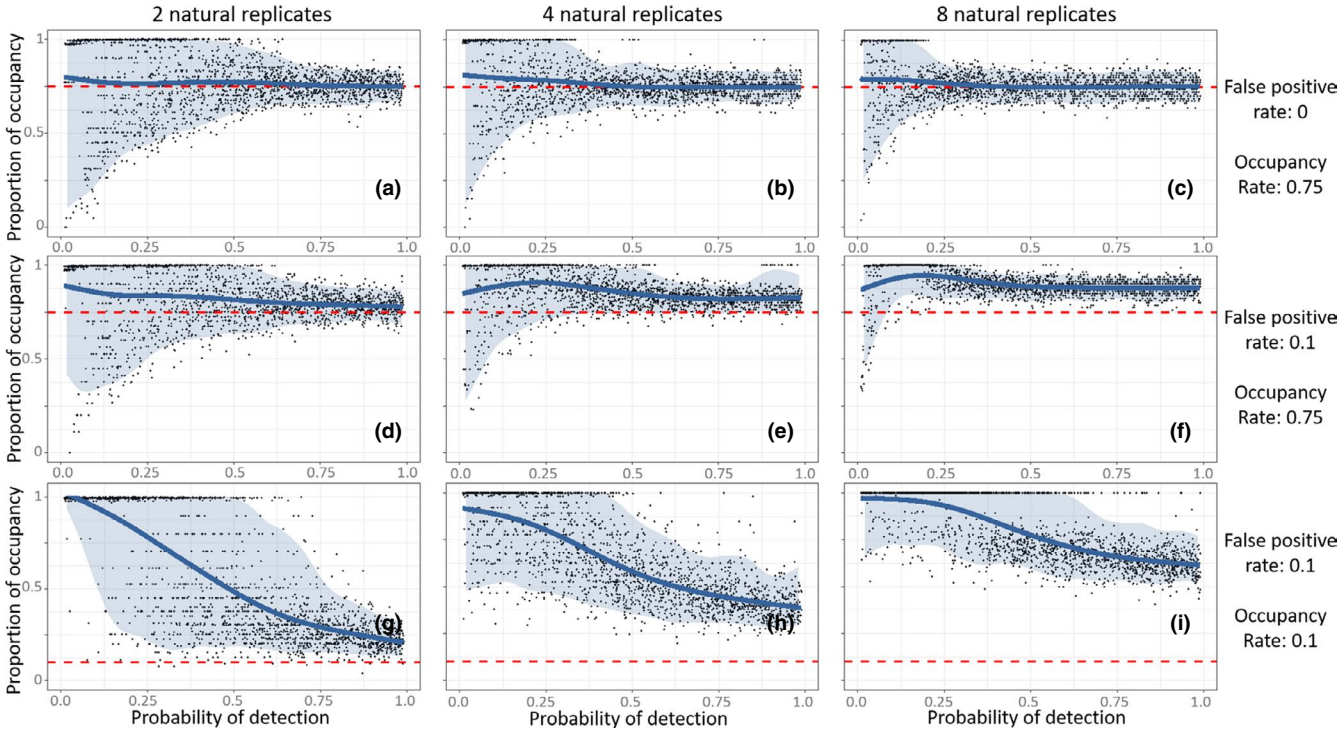


FIGURE 2 The performance of occupancy models varies in response to a number of factors including sampling design (e.g. number of natural replicates), reliability of data collection (false-positive and false-negative rates) and species' occurrence (occupancy rates). Here, we assess interactions between these factors in their impact on the reliability of model predictions. Each of the simulated 2000 points per panel reflects a landscape with 80 sampling sites (see the supplementary information for details). Panels present comparisons of the true rate of occupancy (dotted red line) with the modelled rate of occupancy (blue line, shaded area reflects 95% range of data) across a range of different detection probabilities. Increases in detection probabilities generally improve model accuracy (difference between true and modelled rate of occupancy) and precision (spread of model outcomes, blue shaded area). In contrast, the impact of higher numbers of natural replicates was strongly dependent on the rate of false positives. When false positives were absent, more natural replicates improved model performance (mostly precision; a–c). However, the opposite was true when false positives occurred (d–f), whereas the impact size of this effect increased with decreasing occupancy rates (g–i). The fact that occupancy modelling performs at first glance better (i.e. higher precision) but in fact is hampered by low accuracy highlights the importance of minimizing false-positive rates as far as possible and accounting for them in modelling approaches

error, see Darling & Mahon, 2011) is that it is mathematically challenging to simultaneously account for false positives and false negatives using a single data set (Miller et al., 2011). New approaches, which have been developed (Guillera-Arroita et al., 2017; Miller et al., 2011) and are in the process of being applied (Louvrier et al., 2019), often depend on the use of multiple data sources (Chambert et al., 2015). Specifically, one data set with high credibility (available for a subset of sampling sites), is initially needed to establish true positive and true negative results (Brost et al., 2018). True positives and negatives can then be compared with a second data set (e.g., eDNA survey results) for the subset of sites where both data sets are available. This comparison allows the estimation of false positive rates with high confidence and therefore enables corrections to be applied, accounting for the occurrence of false positives across the entire eDNA data set. Indeed, it should be remembered that detecting an eDNA "signal" does not necessarily imply the truce presence of a species.

Interestingly, the techniques developed to account for false positives have opened up new and exciting possibilities to combine multiple data sources through occupancy modelling approaches (Miller et al., 2019). Recent evaluations have demonstrated that the reliability of predictions can be substantially increased if different data sources are combined—even if these sources vary widely in their reliability (Lahoz-Monfort et al., 2016). Consequently, the combination of traditional and eDNA survey programmes and the use of alternative data sources (e.g. those generated by citizen science programmes) provide ample possibilities to advance current approaches and support better-informed decision-making processes. Indeed, citizen science has the potential to collect large volumes of data over vast areas (Larson et al., 2020). Despite potential bias when covering larger geographical areas, data quality has been shown to be highly improved with limited training and the use of validated and standardized protocols (Larson et al., 2020) as discussed above.

4.3 | Integration of spatial patterns and process-based tools

Another key component that is starting to be integrated into species distribution (Domisch et al., 2019; Pacifici et al., 2017) and eDNA occupancy models (Chen & Ficetola, 2019) is the spatial distribution of species occurrence in their habitat. A number of factors, including spatial autocorrelation of environmental conditions, population distribution patterns and/or eco-geographical factors, result in spatial coupling of species occurrence (Legendre, 1993). Spatial population structures can be accounted for in occupancy models by integrating autoregressive terms (Domisch et al., 2019; Pacifici et al., 2017), which increases or decreases the probability of occupancy at one site, depending on the occupancy of adjacent sites. Autoregressive terms can, for example, help to identify a single positive detection in an area otherwise characterized by the absence of the target species as a false positive. Similarly, they will result in an adjusted likelihood of a negative result being incorrect if it is found among

BOX 2 Spatially explicit models in river networks

Spatial non-independence is a common phenomenon across terrestrial and aquatic habitats. The consideration of spatial dependencies is therefore an important step in the determination of species distributions. Despite the generality of spatial interdependencies, most methodological advances accounting for such effects have been generated in aquatic research. In freshwater ecosystems, upstream environments tend to influence more downstream-located habitats. The resulting directionality and overall nestedness leads to a strong spatial autocorrelation among river reaches (Legendre, 1993) that needs to be accounted for in any spatial model (Dormann et al. 2007). Dormann et al. (2007) provided three main arguments highlighting the necessity to integrate spatial autocorrelation in modelling approaches: (i) species dispersal is distance-related, (ii) the nonlinear relationships between environment and species cannot be modelled as linear, and (iii) the fact that a non-spatial statistical model would fail to account for environmental determinants, which are spatially structured, and whose spatial structuring cascades into the response.

One way to account for spatial autocorrelation in river networks is provided by simultaneous autoregressive (SAR) models. SAR models represent the directed version of conditional autoregressive (CAR) models and incorporate the possibility to apply an asymmetric covariance matrix. More recently, Peterson et al. (2013) have introduced the spatial statistical stream network (SSN) model framework, tailored explicitly towards stream network applications; Hoef et al., 2014, Ver Hoef & Peterson, 2010). Here, the model allows us to accommodate so-called "tail-up" and "tail-down" models, where the former refers to accounting for autocorrelation between flow-connected locations, while the latter also allows spatial autocorrelation between flow-connected and flow-unconnected locations (Peterson et al. 2013).

Models that account for spatial autocorrelation outperform nonspatial models (Domisch et al., 2019; Ver Hoef & Peterson, 2010), but the preprocessing regarding hydrological connectivity and spatial weights in the spatial models requires advanced GIS skills, posing a challenge to the wide application across disciplines in freshwater research.

Occupancy models (see Web Panel 1) account for the detection probability of species, which is crucial especially when modelling species that are difficult to detect (Comte & Grenouillet, 2013). Such models can be extended to spatially explicit occupancy models by incorporating spatial random effects via CAR or SAR component in the model (Chen & Ficetola, 2019; Latimer et al., 2006), and can then be applied to river networks (Domisch et al., 2016, Domisch et al., 2019).

several positive measurements. Applications of autoregressive terms in eDNA occupancy and species distribution models have only recently been demonstrated to improve model performances (Chen & Ficetola, 2019), and represent a promising approach to enhance the reliability of eDNA surveys.

Autoregressive models have also been developed to account for directed effects (e.g., simultaneously autoregressive models; see Box 2). Directed autoregressive terms are certainly powerful tools but they have limitations when dealing with systematic errors (Box 2) such as the downstream transport of eDNA in rivers causing false positive results (Pont et al., 2018). A more bespoke tool to correct for systematic and directed increases in the risk of false positives is the application of process-based models.

Process-based models provide an alternative to statistical approaches and can help account for systematically occurring false positive or false negative results. In contrast to occupancy models, these are based on a mechanistic understanding of the dynamics of eDNA concentrations in the environment. Consequently, process-based models critically rely on the accurate quantification of eDNA concentrations, and their application will be promoted by advancements such as droplet digital PCR increasing measurement precision (e.g., Doi et al., 2015; Orzechowski et al., 2019). We provide here, as an example of a process-based model, the correction of false positive results caused by the downstream transport of eDNA across multiple sites in a hypothetical river network (Figure 3a). Patterns in species distribution and population densities drive the dynamics of eDNA concentration in the water (Figure 3b). When the target species is absent in a river reach, eDNA concentrations decrease logarithmically, but false positive detection is possible due to the downstream transport of eDNA.

To account for such false positives, eDNA export curves can be established to quantify the transport of upstream eDNA to downstream habitats. Such export curves can be deduced from hydrological models and mesocosm eDNA degradation experiments (e.g., Seymour et al., 2018; Song et al., 2017), or simply by measuring *in situ* the downstream transport of introduced eDNA (e.g., Pont et al., 2018). Once established, export curves from upstream sampling points (Figure 3b) can be compared to measured eDNA concentrations of downstream sites. This requires the establishment of upper ($C_{pred.upper}$) and lower ($C_{pred.lower}$) prediction intervals for downstream-transported eDNA concentrations based on

$$C_{pred.upper} = \left(Conc_{0mean} + Conc_{0SEM}\right) * slope_{upper}^{d_{P1-P0}} \quad (1)$$

$$C_{pred.lower} = \left(Conc_{0mean} - Conc_{0SEM}\right) * slope_{lower}^{d_{P1-P0}} \quad (2)$$

where $Conc_{0\ mean}$ and $Conc_{0\ SEM}$ represent the mean and standard error of the mean of the upstream sampling site (*P*0), $slope_{upper}$ and $slope_{lower}$ stand for the upper and lower confidence interval of the slope of the export curve and $d_{P1\text{-}P0}$ is the distance between P0 and the downstream sampling site (*P*1). If the 95% confidence interval of the mean eDNA concentration at *P*1 overlaps with the computed prediction interval (e.g., at point 4 in Figure 3b), false positive risks are inflated, and additional sampling methods should be applied. Alternatively, the output of process-based models can be included as prior information in hierarchical (e.g. occupancy) models that account for false positives, which allows capitalizing on synergies between the two approaches.

5 | OUTLOOK AND FUTURE CHALLENGES

Good field and laboratory practices are essential for minimizing many sources of error associated with the use of eDNA, but even best practices cannot exclude the occurrence of false positive and false negative results. Data processing tools such as occupancy or process-based models provide opportunities to mitigate the impact of many sources of error. Increases in the size of eDNA data sets

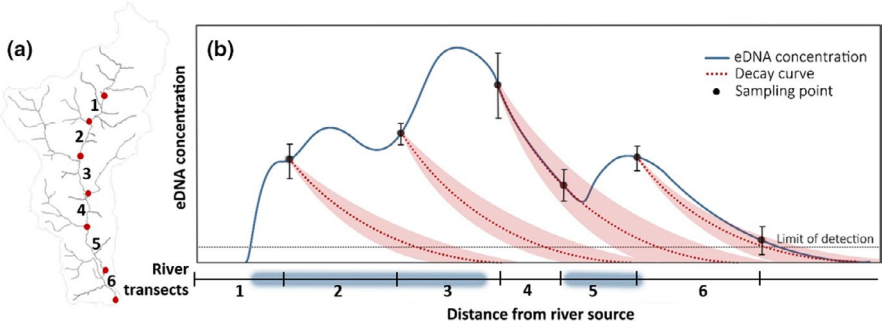


FIGURE 3 Example of the application of a mechanistic model to increase the precision of eDNA-based occupancy predictions. (a) Schematic of river catchment surveyed along a longitudinal transect. River sections are defined as sections between sampling points. (b) Change of eDNA concentration (blue line) along the same longitudinal transect. The transport of eDNA from upstream river transects can lead to the presence of eDNA in the absence of the target species in the downstream transect (presence of target species indicated as blue-shaded areas on the *x*-axis). Without corrections, this can lead to false-positive results (e.g. in transects 4 and 6). A mechanistic model based on eDNA decay rates, water flow velocity and changes in water flow can allow the establishment of eDNA decomposition curves (red dotted lines) for each sampling point (black dots). If downstream measurements of eDNA concentrations overlap with the confidence intervals of the upstream eDNA decomposition curve, the presence of the target species in a given river transect is questionable and should be reinvestigated with alternative methods (examples are transects 4 and 6)

due to improved cost-efficiency and further technical advances will substantially help to further improve the power of these data processing tools. However, eDNA-based monitoring should not be seen in isolation or as a replacement for traditional survey approaches. The true potential of eDNA-based methods can only be capitalized on when they are combined with other sampling data and jointly integrated with data processing tools (Miller et al., 2019; Pacifici et al., 2017). Consequently, an important future challenge will be the coordination and scaling of different assessments, such as traditional sampling methods, eDNA-based methods and citizen science campaigns. Only together can these approaches raise the necessary public awareness and provide the reliable baseline data required to meet future challenges in conservation and ecosystem management.

ACKNOWLEDGEMENTS

The study was funded by the Global Challenges Research Fund, UK, to M.S. and the Leibniz Competition to S.D. (J45/2018).

DATA AVAILABILITY STATEMENT

Data are provided in supplementary information.

ORCID

Alfred Burian https://orcid.org/0000-0002-4928-0897
Quentin Mauvisseau https://orcid.org/0000-0003-1215-2987
Michael Sweet https://orcid.org/0000-0003-4983-8333

REFERENCES

Bailey, L. L., MacKenzie, D. I., & Nichols, J. D. (2014). Advances and applications of occupancy models. *Methods in Ecology and Evolution*, *5*, 1269–1279.

Brost, B. M., Mosher, B. A., & Davenport, K. A. (2018). A model-based solution for observational errors in laboratory studies. *Molecular Ecology Resources*, *18*, 580–589.

Buxton, A. S., Groombridge, J. J., Zakaria, N. B., & Griffiths, R. A. (2017). Seasonal variation in environmental DNA in relation to population size and environmental factors. *Scientific Reports*, *7*, 46294.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*, 581–583.

Chambert, T., Miller, D. A. W., & Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, *96*, 332–339.

Chen, W., & Ficetola, G. F. (2019). Conditionally autoregressive models improve occupancy analyses of autocorrelated data: An example with environmental DNA. *Molecular Ecology Resources*, *19*, 163–175.

Comte L., Grenouillet G. (2013). Species distribution modelling and imperfect detection: comparing occupancy versus consensus methods. *Diversity and Distributions*, *19*(8), 996–1007. https://doi.org/10.1111/ddi.12078

Cristescu, M. E., & Hebert, P. D. N. (2018). Uses and misuses of environmental DNA in biodiversity science and conservation. *Annual Review of Ecology Evolution and Systematics*, *49*(1), 209–230.

Darling, J. A., & Mahon, A. R. (2011). From molecules to management: Adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environmental Research*, *111*, 978–988.

Deiner, K., Lopez, J., Bourne, S., Holman, L., Seymour, M., Grey, E. K., Lacoursière, A., Li, Y., Renshaw, M. A., Pfrender, M. E., Rius, M., Bernatchez, L., & Lodge, D. M. (2018). Optimising the detection of marine taxonomic richness using environmental DNA metabarcoding: The effects of filter material, pore size and extraction method. *Metabarcoding Metagenomics*, *2*, e28963.

Doi, H., Fukaya, K., Oka, S.-I., Sato, K., Kondoh, M., & Miya, M. (2019). Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model. *Scientific Reports*, *9*, 3581. https://doi.org/10.1038/s41598-019-40233-1

Doi, H., Takahara, T., Minamoto, T., Matsuhashi, S., Uchii, K., & Yamanaka, H. (2015). Droplet digital polymerase chain reaction (PCR) outperforms real-time PCR in the detection of environmental DNA from an invasive fish species. *Environmental Science and Technology*, *49*, 5601–5608.

Domisch, S., Friedrichs, M., Hein, T., Borgwardt, F., Wetzig, A., Jähnig, S. C., & Langhans, S. D. (2019). Spatially explicit species distribution models: A missed opportunity in conservation planning?. *Diversity and Distributions*, *25*, 758–769.

Domisch S., Wilson A. M., Jetz W. (2016). Model-based integration of observed and expert-based information for assessing the geographic and environmental distribution of freshwater species. *Ecography*, *39*(11), 1078–1088. https://doi.org/10.1111/ecog.01925

Dorazio, R. M., & Erickson, R. A. (2018). ednaoccupancy: An R package for multi-scale occupancy modeling of environmental DNA data. *Molecular Ecology Resources*, *18*, 368–380.

Dormann C. F., McPherson J. M., Araújo M. B., Bivand R., Bolliger J., Carl G., Davies R. G., Hirzel A., Jetz W., Daniel Kissling W., Kühn I., Ohlemüller R., Peres-Neto P. R., Reineking B., Schröder B., Schurr F. M., Wilson R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, *30*(5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Evans, N. T., Shirey, P. D., Wieringa, J. G., Mahon, A. R., & Lamberti, G. A. (2017). Comparative cost and effort of fish distribution detection via environmental DNA analysis and electrofishing. *Fisheries*, *42*, 90–99.

Ferguson, P. F. B., Conroy, M. J., & Hepinstall-Cymerman, J. (2015). Occupancy models for data with false positive and false negative errors and heterogeneity across sites and surveys. *Methods in Ecology and Evolution*, *6*, 1395–1406.

Ficetola, G. F., Pansu, J., Bonin, A. et al (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*, 543–556.

Ficetola, G. F., Taberlet, P., & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Molecular Ecology Resources*, *16*, 604–607.

Goldberg, C. S., Strickler, K. M., & Fremier, A. K. (2018). Degradation and dispersion limit environmental DNA detection of rare amphibians in wetlands: Increasing efficacy of sampling designs. *Science of the Total Environment*, *633*, 695–703.

Goldberg, C. S., Turner, C. R., Deiner, K. et al (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*, 1299–1307.

Griffin, J. E., Matechou, E., Buxton, A. S., Bormpoudakis, D., & Griffiths, R. A. (2020). Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(2), 377–392. https://doi.org/10.1111/rssc.12390

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: Advances, challenges and opportunities. *Ecography*, *40*, 281–295.

Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R. et al (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution*, *8*, 1081–1091.

Hansen, B. K., Bekkevold, D., Clausen, L. W., & Nielsen, E. E. (2018). The sceptical optimist: challenges and perspectives for the application

of environmental DNA in marine fisheries. *Fish and Fisheries*, *19*(5), 751–768. https://doi.org/10.1111/faf.12286

Harper, L. R., Lawson Handley, L., Hahn, 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*). *Ecology and Evolution*, *8*, 6330–6341.

Hoef J. M. V., Peterson E. E., Clifford D., Shah R. (2014). SSN: AnRPackage for Spatial Statistical Modeling on Stream Networks. *Journal of Statistical Software*, *56*(3), https://doi.org/10.18637/jss.v056.i03

IPBES. (2019). Media Release: Nature's Dangerous Decline 'Unprecedented'; Species Extinction Rates 'Accelerating'. Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.

Klymus, K. E., Merkes, C. M., Allison, M. J., Goldberg, C. S., Helbing, C. C., Hunter, M. E., Jackson, C. A., Lance, R. F., Mangan, A. M., Monroe, E. M., Piaggio, A. J., Stokdyk, J. P., Wilson, C. C., & Richter, C. A. (2020). Reporting the limits of detection and quantification for environmental DNA assays. *Environmental DNA*, *2*(3), 271–282. https://doi.org/10.1002/edn3.29

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

Larson, E., Graham, B., Achury, R., Coon, J., Daniels, M., Gambrell, D., Jonasen, K., King, G., LaRacuente, N., Perrin-Stowe, T., Reed, E., Rice, C., Ruzi, S., Thairu, M., Wilson, J., & Suarez, A. (2020). From eDNA to citizen science: emerging tools for the early detection of invasive species. *Frontiers in Ecology and the Environment*, *18*(4), 194–202. https://doi.org/10.1002/fee.2162

Latimer A. M., Wu S., Gelfand A. E., Silander Jr J. A. (2006). Building Statistical Models To Analyze Species Distributions. *Ecological Applications*, *16*(1), 33–50. https://doi.org/10.1890/04-0609

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology*, *74*, 1659–1673.

Li, J., Lawson Handley, L.-J., Read, D. S., & Hänfling, B. (2018). The effect of filtration method on the efficiency of environmental DNA capture and quantification via metabarcoding. *Molecular Ecology Resources*, *18*, 1102–1114.

Louvrier, J., Molinari-Jobin, A., Kéry, M., Chambert, T., Miller, D., Zimmermann, F., Marboutin, E., Molinari, P., Müeller, O., Černe, R., & Gimenez, O. (2019). Use of ambiguous detections to improve estimates from species distribution models. *Conservation Biology*, *33*, 185–195.

Mackenzie, D. I., & Royle, J. A. (2005). Designing occupancy studies: General advice and allocating survey effort: designing occupancy studies. *Journal of Applied Ecology*, *42*, 1105–1114.

Mauvisseau, Q., Burian, A., Gibson, C., Brys, R., Ramsay, A., & Sweet, M. (2019). Influence of accuracy, repeatability and detection probability in the reliability of species-specific eDNA based approaches. *Scientific Reports*, *9*, 580. https://doi.org/10.1038/s41598-018-37001-y

McClenaghan, B., Compson, Z. G., & Hajibabaei, M. (2020). Validating metabarcoding-based biodiversity assessments with multi-species occupancy models: A case study using coastal marine eDNA. *PLoS One*, *15*(3), e0224119.

Merkes, C. M., McCalla, S. G., Jensen, N. R., Gaikowski, M. P., & Amberg, J. J. (2014). Persistence of DNA in carcasses, slime and avian feces may affect interpretation of environmental DNA data. *PLoS One*, *9*, e113346.

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, *92*, 1422–1428.

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*, 22–37.

Orzechowski, S. C. M., Frederick, P. C., Dorazio, R. M., & Hunter, M. E. (2019). Environmental DNA sampling reveals high occupancy rates of invasive Burmese pythons at wading bird breeding aggregations in the central Everglades. *PLoS One*, *14*, e0213943.

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., & Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion*. *Ecology*, *98*, 840–850.

Peterson E. E., Ver Hoef J. M., Isaak D. J., Falke J. A., Fortin M. J., Jordan C. E., McNyset K., Monestiez P., Ruesch A. S., Sengupta A., Som N., Steel E. A., Theobald D. M., Torgersen C. E., Wenger S. J. (2013). Modelling dendritic ecological networks in space: an integrated network perspective. *Ecology Letters*, *16*, (5), 707–719. https://doi.org/10.1111/ele.12084

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R., Parris, K., Vesk, P., & Mccarthy, M. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, *5*, 397–406.

Pont, D., Rocle, M., Valentini, A., Civade, R., Jean, P., Maire, A., Roset, N., Schabuss, M., Zornig, H., & Dejean, T. (2018). Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. *Scientific Reports*, *8*–10361. https://doi.org/10.1038/s41598-018-28424-8

Rota, C. T., Fletcher, R. J. Jr, Dorazio, R. M., & Betts, M. G. (2009). Occupancy estimation and the closure assumption. *Journal of Applied Ecology*, *46*, 1173–1181.

Ruiz-Gutierrez, V., Hooten, M. B., & Campbell Grant, E. H. (2016). Uncertainty in biological monitoring: A framework for data collection and analysis to account for multiple sources of sampling bias. *Methods in Ecology and Evolution*, *7*, 900–909.

Sales N. G., Wangensteen O. S., Carvalho D. C., Deiner K., Præbel K., Coscia I., McDevitt A. D., Mariani S. (2021). Space-time dynamics in monitoring neotropical fish communities using eDNA metabarcoding. *Science of The Total Environment*, *754*, 142096. https://doi.org/10.1016/j.scitotenv.2020.142096

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*, 1289–1303.

Seymour, M., Durance, I., Cosby, B. J., Ransom-Jones, E., Deiner, K., Ormerod, S. J., Colbourne, J. K., Wilgar, G., Carvalho, G. R., de Bruyn, M., Edwards, F., Emmett, B. A., Bik, H. M., & Creer, S. (2018). Acidity promotes degradation of multi-species environmental DNA in lotic mesocosms. *Communications Biology*, *1*–4. https://doi.org/10.1038/s42003-017-0005-3

Song, J. W., Small, M. J., & Casman, E. A. (2017). Making sense of the noise: The effect of hydrology on silver carp eDNA detection in the Chicago area waterway system. *Science of the Total Environment*, *605–606*, 713–720.

Spens, J., Evans, A. R., Halfmaerten, D., Knudsen, S. W., Sengupta, M. E., Mak, S. S. T., Sigsgaard, E. E., & Hellström, M. (2017). Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. *Methods in Ecology and Evolution*, *8*, 635–645.

Troth, C., Sweet, M., Nightingale, J., & Burian, A. (2021). Seasonality, DNA degradation and spatial heterogeneity as drivers of eDNA detection dynamics. *Science of the Total Environment*, *768*, 144466.

Ver Hoef J. M., Peterson E. E. (2010). A Moving Average Approach for Spatial Statistical Models of Stream Networks. *Journal of the American Statistical Association*, *105*(489), 6–18. http://dx.doi.org/10.1198/jasa.2009.ap08248

Wilcox, T. M., McKelvey, K. S., Young, M. K., Jane, S. F., Lowe, W. H., Whiteley, A. R., & Schwartz, M. K. (2013). Robust detection of rare species using environmental DNA: The importance of primer specificity. *PLoS One*, *8*, e59520.

Wood, S. A., Biessy, L., Latchford, J. L., Zaiko, A., von Ammon, U., Audrezet, F., Cristescu, M. E., & Pochon, X. (2019). Release and degradation of environmental DNA and RNA in a marine system. *Science of the Total Environment*, *704*, 135314.

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*, 1857–1862.

Zizka, V. M. A., Elbrecht, V., Macher, J., & Leese, F. (2019). Assessing the influence of sample tagging and library preparation on DNA metabarcoding. *Molecular Ecology Resources*, *19*, 893–899.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Burian A, Mauvisseau Q, Bulling M, Domisch S, Qian S, Sweet M. Improving the reliability of eDNA data interpretation. *Mol Ecol Resour*. 2021;00:1–12. https://doi.org/10.1111/1755-0998.13367

APPENDIX 1

WEB PANEL 1

Introduction to occupancy modelling

An occupancy model (MacKensie et al., 2018) at its fundamental level is a mixed model, a mixture of two probability distributions to describe the two sources of uncertainty during a survey. One is the sampling variation and the other is the imperfect detection. In the simplest case, a site is repeatedly visited to detect the presence of a target species. By chance, the target species may not be present during the time of a specific visit. As a result, the number of detections x from a total number of visits n is a random variable and is usually modelled by the binomial distribution. In almost all occupancy surveys, our method of detection is rarely perfect. We may fail to detect the target species when it is present. The imperfect detection process is modelled by a Bernoulli distribution (a special case of the binomial distribution when $n = 1$). When we conduct a survey, we observed the number of times (x) we detect the target in n visits. Because of imperfect detection, we cannot definitely tell that an observed 0 is because the target is absent or because of detection failure. However, mathematically, we can describe the data-generating process as a result of the two separate random processes.

The number of presences (x) is modelled by the binomial distribution:

$$x \sim bin(\theta, n) \tag{A1}$$

where θ is the probability of detecting the target species. The probability θ is the product of the probability of being present (ψ) and the probability of detecting the target when it is present (p). The probability of being present ψ is the occupancy probability and the probability of detection is a conditional probability characterizing the survey method. Both probabilities are of interest. However, the data we have (x, n) has only information for $\theta = \psi \times p$. That is, ψ and p are numerically unidentifiable in the simplest case. Additional information is needed for separating ψ from p.

A typical occupancy study is designed to provide information to separately estimate ψ and p. For example, the ad hoc two-step approach developed by Geissler and Fuller (1987) is designed to estimate p first and then used the observed x to estimate θ. The process of estimating p requires repeated sampling of the same sites and counting the number of times the target species is detected and the total number of visits after the first detection (when the presence of the target species is confirmed). The detection probability is then approximated by the number of detections divided by the number of visits. Other survey designs are aimed at using covariates to provide the necessary information to better separate the two probabilities. These designs do not always work well under the classical statistics framework, where the maximum likelihood estimator is usually used. The underlying numerical identification problem is always lurking, in addition to the inherent positive correlation between the detection probability and the occupation probability. That is, the more abundant the target species is, the easier it is to detect them.

Bayesian analysis of occupancy models

The arrival of the MCMC method (Gelfand et al., 1990; Gelfand and Smith, 1990), especially its computer implementation in software packages such as WINBUGS, JAGS and now STAN made the computation under the Bayesian seemingly straightforward. For example, the simplest model can be expressed by introducing a latent variable z, indicating whether a detection occurs:

$$\begin{aligned} x &\sim bin(z\psi, n) \\ z &\sim bern(p) \end{aligned} \tag{A2}$$

Implementation of this model under MCMC (using WINBUG, JAGS or STAN) is straightforward. Typically, vague or flat priors are used for model parameters. Such a model will often run using MCMC. However, the numerical identifiability issue reflected in the highly correlated joint posterior distribution of ψ and p remains. When marginal posterior distributions of ψ and p are reported (the default output format of almost all MCMC software packages), we are often not aware of the high correlation between ψ and p; rather, we observe either widespread marginal distributions or, more

probably, highly skewed marginal distributions concentrated near 0 or 1. The problem of misrepresenting the correlated joint posterior distribution by the posterior marginal distributions is quite common (Qian, 2012). Without additional information (e.g., a proper informative joint prior), these numerical issues will always be present. Consequently, the key to a successful Bayesian occupancy model lies in the development of a proper joint prior distribution of the two probabilities.

When using eDNA for occupancy modelling, we face not only the imperfect detection probability (false negative), but also false positive. Let p_p and p_n be the probability of a false positive and false negative, respectively (note that $p = 1 - p_n$). The observed number of presence x is still a binomial random variable, but the probability of observing a presence (positive eDNA) is now $\theta = \psi(1 - p_n) + (1 - \psi)p_p$. Without proper (informative) priors for p_p and p_n, occupancy modelling will always be numerically unstable.

REFERENCES

Geissler, P. H., & Fuller, M. R. (1987). Estimation of the proportion of area occupied by an animal species. In: *Proceedings of the Section on Survey Research Methods* (pp. 533–538). American Statistical Association.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, *85*, 972–985.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., & Hines, J. E. (2018). *Occupancy Estimation and Modeling: Inferring partterns and dynamics of species occurrence* (2nd Ed). : Academic Press.

Qian, S. S. (2012). On model coefficient estimation using Markov chain Monte Carlo simulations: A potential problem and the solution. *Ecological Modelling*, *32*(6), 297–304.