


Integrating freshwater biodiversity data sources: Key challenges and opportunities

Susan G. Jarvis¹  | Eleanor B. Mackay¹ | Hannah A. Risser¹  | Heidrun Feuchtmayr¹ | Matthew Fry² | Nick J. B. Isaac² | Stephen J. Thackeray¹ | Peter A. Henrys¹

¹UK Centre for Ecology & Hydrology, Lancaster Environment Centre, Lancaster, UK

²UK Centre for Ecology & Hydrology, Wallingford, UK

Correspondence

Susan G. Jarvis, UK Centre for Ecology & Hydrology, Lancaster Environment Centre, Lancaster LA1 4AP, UK.

Email: susjar@ceh.ac.uk

Funding information

Department for Environment, Food and Rural Affairs, UK Government; Natural Environment Research Council

Abstract

1. In order to better quantify spatial and temporal patterns in freshwater biodiversity, and potential underlying drivers of change, we must utilise the increasingly broad range of data available on freshwater ecosystems. Statistical advances in the field of integrated modelling provide new opportunities to further our understanding through the combined and simultaneous analysis of these diverse datasets.
2. We briefly introduce integrated modelling in the context of freshwater biodiversity and outline the key steps involved in its implementation, from data collection to analysis. We highlight both opportunities and challenges for the application of integrated approaches.
3. To illustrate the potential for integrated models to improve our understanding of freshwater biodiversity compared to standard approaches, we combine two datasets collected using different methods to model the distribution of *Agabus* water beetles in England. The integrated model had greater power to detect covariate effects on *Agabus* distribution, and reduced parameter uncertainty compared with analysis using only a single dataset.
4. We show that integrated methods have the potential to increase our understanding of freshwater systems and enable us to make full use of the diversity of freshwater data available.

KEYWORDS

citizen science, data access, integrated model, statistics, water beetle

1 | INTRODUCTION

High rates of freshwater biodiversity change and decline, and ecological degradation, are well-documented at a global scale; a result of the combined action of several globally pervasive pressures, including pollution, habitat alteration, species introduction, over-exploitation and climate change (Jenny et al., 2020; Revenga et al., 2005; Tickner et al., 2020). Our ability to accurately quantify

large-scale spatiotemporal patterns of change and understand underlying drivers depends upon robust and representative data. These are challenging to collect in highly heterogeneous and ecologically complex environments, especially given the resource limitations typically faced by researchers and agencies. While this is a significant cause for concern, we now have access to a widening array of monitoring approaches that allow us to quantify and attribute freshwater ecosystem status, change and underlying drivers (Thackeray

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Freshwater Biology* published by John Wiley & Sons Ltd.

& Hampton, 2020). For freshwater ecosystems, the range of available monitoring methods now includes such diverse approaches as sensor networks, ecoacoustics, eDNA, remote sensing, text mining, traditional *in situ* sampling, museum records, indigenous knowledge and rapidly-growing citizen science initiatives (Harper et al., 2019; Jarić et al., 2020; Linke et al., 2018; Metcalfe et al., 2022; Palmer et al., 2015; Pellerin et al., 2016; Revenga et al., 2005, and references therein). These different monitoring approaches will vary in their power to detect change and driver effects at different spatio-temporal scales and levels of biological organisation.

The purpose of monitoring also varies depending on the study focus and research questions; therefore, data are frequently collected using different survey designs. Structured monitoring, such as that often required to meet statutory requirements under the EU Water Framework Directive, with robust underlying design, repeatable protocols, repeated observations over time and quality assurance processes are typically seen as the gold standard for biodiversity monitoring as they provide the greatest control over the data collection process (Buckland & Johnston, 2017). However, resource limitations and logistical challenges often mean that structured monitoring is limited in the number of locations that can be surveyed. Unstructured or opportunistic data collection – data collected without a strict design or protocol – has emerged as a means of widening spatial coverage and temporal resolution of the natural environment. Wider use of unstructured data has now been made possible as a result of the development of statistical techniques and computational resources for analysing large volumes of incomplete or biased data (Isaac et al., 2014).

The current proliferation and diversification of data sources provides opportunities to enhance our fundamental understanding of the behaviour of freshwater ecosystems, providing evidence to guide management and restoration efforts. However, to fully capitalise on this potential and unlock the “complementarities” that exist among methods (Thackeray & Hampton, 2020), we need the capability to bring different sources of data together to answer questions about how freshwater biodiversity is changing, and why. Combining data collected using different methods and with different designs provides a challenge for data analysts, particularly when data are very different from each other (e.g., eDNA and *in situ* observations). To address this challenge, researchers have developed integrated modelling approaches which allow data from multiple sources to be combined within the same analytical workflows, allowing different datasets to be used simultaneously to assess ecological status and trends (DeWan & Zipkin, 2010; Fletcher Jr et al., 2019; Isaac et al., 2020; Miller et al., 2019). Integrated models can combine data collected using different methodologies and designs by accounting for differences between data sources within the model structure.

Bringing multiple datasets together to address environmental problems has a number of advantages beyond simply increasing the pool of data that can be used:

1. Well-constructed integrated models inherit the strengths of the datasets that contribute to them, and also counter their

weaknesses, for example by utilising the large sample size offered by opportunistic citizen science data and the robust design of a systematic survey.

2. Integration of data with different properties makes it possible to capture processes operating at different spatial and temporal scales (Ryo et al., 2019; Zipkin et al., 2021), or to estimate parameters that would not be identifiable using a single dataset.
3. Integrated models can allow better estimation of potential environmental driver effects than using single datasets alone through increased coverage of environmental gradients (Bowler et al., 2019).
4. Where source datasets present conflicting signals about biodiversity change, an integrated framework has the potential to reveal where the key uncertainties lie, to target future data collection.

Integrated modelling is seeing increased uptake in the ecological literature, but the vast majority of existing applications are from terrestrial or marine environments (Bowler et al., 2019; Martino et al., 2021; Zulian et al., 2021), and there has been limited application within the freshwater domain despite there being a strong case to do so (Bishop et al., 2021). The aim of this paper is to increase awareness of integrated analysis within the freshwater ecological community via a brief introduction to integrated modelling and the existing literature, outlining the steps required to conduct a successful integrated analysis, and presenting an example of integrated modelling in a freshwater context through a case study of British water beetles.

1.1 | Introduction to integrated modelling

Recent developments in applied statistics mean that it is now possible to build statistical models that combine very different types of ecological data within a single model, such as visual surveys with acoustic monitoring, or field measurements with remote sensing (Henry & Jarvis, 2019; Zulian et al., 2021). To achieve this, there is an underlying assumption that there exists at least one shared ecological state, process or parameter common to different datasets, albeit measured in different ways (Figure 1). Integrated models allow for differences in the way observations are made by adding sub-models that account for different observation processes (Figure 1; Isaac et al., 2020). Differences between data collection methods and survey designs can be represented using additional covariates or a hierarchical error structure that allows key parameters within the model to differ across datasets (Moriarty et al., 2020; Piepho & Ogutu, 2002).

2 | STEPS IN AN INTEGRATED ANALYSIS

Given the potential advantages of integrated modelling in addressing questions in freshwater ecology we aim to provide a helpful

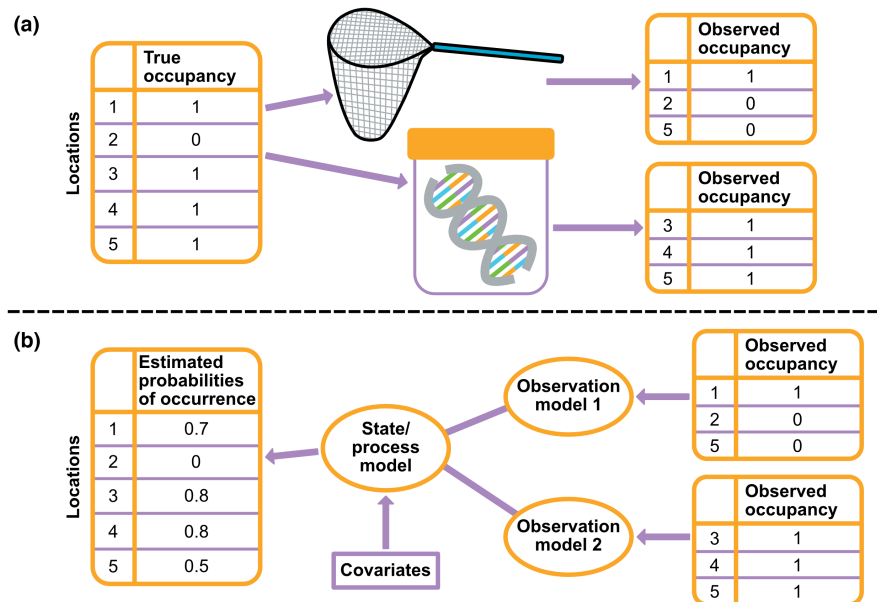


FIGURE 1 Hypothetical example of integrated modelling. (a) Two datasets reflecting the same ecological state (whether or not a location is occupied by a species) but observed in different ways, one through net sampling and one through an eDNA approach. Both approaches give information on when the species was observed, but the eDNA data do not have information on locations which were sampled but where a sequence was not found (i.e., no zeroes). Both observation methods have detection probabilities lower than one, that is they will not detect all individuals or sequences present (e.g., location 5 in the net survey records a 0 even though the site is actually occupied) and only location 5 has data from both survey types. However, both datasets relate back to the same underlying ecological state (true occupancy per location), even though they are observed in different ways. (b) How integrated models work: both datasets are used to estimate the probability of each location being occupied by the species. To do this we model both the ecological state (whether or not the location is occupied) and the observation processes (e.g., net or eDNA method), allowing each dataset to be observed in a different way through the addition of observation sub-models. We also could include covariates; in the figure these feed into the state model only but covariates might be used in the observation sub-models if they are likely to affect detection.

introduction by highlighting four key steps in a successful integrated modelling application: collecting (and storing), mobilising, exploring and analysing data. The first two steps are generally done on each dataset separately, whereas exploration and analysis are conducted on multiple datasets simultaneously. Although the analysis step may seem most obvious when we consider an introduction to integrated models, we wish to highlight elements of all four steps that can influence how successfully datasets can be analysed jointly. We briefly discuss each step, highlighting both opportunities and challenges.

2.1 | Data collection and storage

A very broad range of freshwater biodiversity data can be potentially suitable for integrated analyses, originating from a wide range of sampling methods and technologies (Kakouei et al., 2021; Read et al., 2014; Rusak et al., 2018). A key consideration for integrating any data is whether there is consistency in the way data are collected and stored within individual datasets that are considered for integrated modelling. Integrated models aim to quantify among-dataset differences owing to variation in the way data are observed, and inconsistencies within a dataset can confound quantification of among-dataset variation. Issues can occur when data are recorded in an inconsistent manner; for example, using

surveyors with different levels of identification skill and experience, collecting different volumes of water, using different sampling devices or using different species naming conventions.

Electronic data capture (e.g., using smartphone apps like Bloomin' Algae for cyanobacterial blooms, iRecord for opportunistic species observations or custom apps; Reaney et al., 2019) can reduce inconsistencies by embedding rules within the software design that increase within-dataset consistency (Murphy & Weatherby, 2008). Modern data capture technologies enable data collected in the field or laboratory to be uploaded manually or automatically onto an underlying database architecture on a central server for storage (Nowak et al., 2020, <http://www.indicia.org.uk/>). Such systems offer many advantages to users and analysts including streamlining the lag between data collection and analysis-ready data, providing a centralised system for record validation, and minimising the likelihood of incurring any errors or inconsistencies.

2.2 | Data access & mobilisation

In order to maximise the potential of integrated modelling we may want to use data collected by other researchers. To do this requires a modeller being able to find, access, understand and evaluate existing data, and requires a commitment from data collectors and holders to

make their data and metadata available to others in a findable way. Data access has been recognised as a key priority for safeguarding global freshwater biodiversity (van Rees et al., 2021).

A valuable resource for freshwater biodiversity is the Freshwater Information Platform (Schmidt-Kloiber et al., 2019) which provides access to data and, more importantly, metadata (data about data) which enables researchers to explore available datasets easily. Access to good metadata is essential to ensure that any data acquired are used appropriately and is a key element of the FAIR (Findable, Accessible, Interoperable and Reusable) data principles (Wilkinson et al., 2016) which, if followed, support data reuse. Aggregation platforms such as the Global Biodiversity Information Facility and National Biodiversity Network are useful resources for species occurrence records, but often lack metadata required to enable data reuse (Turner et al., 2023). Data access is a particular challenge for historical datasets which are rarely available online, yet may provide important insights into long-term ecological processes.

Integrated modelling also is facilitated by open access data (i.e., data available to all). Open data provide both opportunities and challenges to those working with biodiversity data. Open access principles unlock biodiversity data which may greatly improve our ability to understand ecological dynamics and pressures in freshwater systems. However, publishing open access data is time-consuming for researchers and can be seen as risk to researchers who must open their datasets for others to publish on and potentially misuse (Mills et al., 2015; Reichman et al., 2011), or a risk to species conservation through exposing locations of sensitive taxa (Tulloch et al., 2018).

2.3 | Data exploration

Once potentially suitable datasets for an integrated modelling application have been identified and accessed, it is important to evaluate the similarities and differences amongst them before building an integrated model. The aims of this data exploration are two-fold: to assess whether integration is a sensible aim and, if so, to identify the key differences that would need to be accounted for in the observation sub-models (Figure 1). In our experience, data exploration is often the most important of the four steps and likely to take the most time in any integrated modelling application.

Properties of any biodiversity monitoring dataset include the taxonomic coverage (i.e., which organisms are included), the taxonomic resolution to which the specimens are identified, and the ecological currency in which data are recorded (e.g., counts vs. presence-absence vs presence-only). Datasets might differ in spatial extent and resolution (e.g., national vs. regional, surface waters vs. depth-resolved) as well as temporal extent and resolution (e.g., time series length and frequency of sampling). Locations may be selected at random or chosen by the surveyor and datasets will often have different sampling protocols, including the time spent collecting data and what equipment is used. Datasets also may vary hugely in the number of observations

taken. Another consideration is whether the observations are subject to any kind of quality assurance, for example to ensure that organisms have been correctly identified and to determine error rates.

When multiple datasets exist ostensibly representing the same environmental phenomena, it can be tempting to consider and compare the quality of the data from each source based on perceived characteristics. For example, data collected as part of a professional scheme may be considered higher quality than data from a voluntary scheme; however, it is important to note that there is not necessarily a clear distinction in quality between schemes using these different groups. There are examples of schemes with a systematic design reliant on volunteers (e.g., Anglers' Riverfly Monitoring Initiative; Brooks et al., 2019) that produce high-quality data. Therefore, it is usually not helpful to think in terms of high- versus low-quality datasets *per se* but rather in terms of the more objective aspects of quality such as the spatial extent, sample size, resolution and design aspects. When data have quality assurance information available, they can be used to assess error rates (e.g., false absences or presences) which may guide choices on dataset use.

Datasets with large differences in their properties may be difficult to combine sensibly in an integrated analysis. This is because, in such a situation, the variance in data resulting from different monitoring scheme properties swamps any shared signals generated by underlying ecological processes. As the area of integrated modelling is still fairly new, there are no clear guidelines as to when data should not be integrated (Simmonds et al., 2020), but there should be sufficient similarity between the data that the assumption of a shared underlying ecological state, process or parameter can be justifiably made. Exploration of the data through graphical means is often informative to make this decision; for example, identifying whether datasets show reasonably similar responses to key, shared environmental gradients. Suitability for integration also will depend on the ecological question at hand and the type of analysis required; for example, if not all datasets capture information on important covariates then it may not be sensible to integrate them. Although data integration does not require large volumes of data, combining datasets of very different sizes may present problems as larger datasets can outweigh the information present in smaller datasets (Fletcher Jr et al., 2019).

In cases where datasets are collected in a sufficiently similar manner to attempt an integrated analysis and it is reasonable to assume data reflect similar or related ecological patterns, the dataset properties will determine the structure of the integrated model needed. To demonstrate these concepts, we consider two UK freshwater invertebrate datasets that monitor water beetles, and present key dataset properties that may influence our ability to conduct an integrated spatial analysis as outlined in Table 1. One dataset is the Environment Agency BIOSYS scheme for bio-indicators of riverine water quality and the other is the world's oldest volunteer-run biological recording scheme, originally known as the Balfour-Browne Club (Balfour-Browne Club, 2020; Foster, 2015).

TABLE 1 Key features of contrasting types of national datasets of UK freshwater invertebrates.

Dataset properties	BIOSYS River macroinvertebrate surveys (BIOSYS)	Water beetle surveys from Britain and Ireland (WBS)
Taxonomic coverage	All aquatic invertebrates	Water beetles, including Sphaeriidae, Gyrinidae, Haliplidae, Noteridae, Paelobiidae, Dytiscidae, Helophoridae, Georissidae, Hydrochidae, Spercheidae, Hydrophilidae, Hydraenidae, Elmidae, Dryopidae, Limnichidae, Heteroceridae, Psephenidae and some other fresh water-associated Coleoptera
Taxonomic resolution	Variable over time, most recent samples to species level except Diptera, Oligochaeta and Bivalvia	Species level
Ecological currency	Counts	Presence-only
Spatial extent	England	UK and most of Ireland
Habitat coverage	Rivers	Rivers, lakes, ponds
Temporal extent	1965 to present day	1904 to present day
Sampling design	Representative, spatially balanced Repeat visits to locations	Volunteer site-selection; opportunistic recording
Sampling protocol	Kick samples and dredge samples	Various, not recorded
Quality assurance	Specimens preserved for verification	Specimens not always preserved but records validated and verified by taxon experts

In the following sections we elaborate on some of these dataset features and the challenges that may arise when considering an integrated analysis.

2.3.1 | Ecological currency

The ecological currency refers to the type of ecological measurement made. The three most common currencies are counts of individuals, presence-absence and presence-only data. Counts contain more information than simply recording presence, yet abundance data are often costly and time-consuming to assemble. Recording the presence of a species is usually more straightforward and presence data can be of two forms. Presence-absence data include non-detections (sampling occasions where the species is not seen) so that some idea of sampling effort and taxonomic focus can be extracted from the survey information. With presence-only data no information is collected when a species was not seen. This difference makes presence-absence data more information-rich, although presence-only data can still be hugely important for assessing biodiversity patterns and trends (Elliott et al., 2015; Huang & Frimpong, 2016). Integrated models enable datasets of different currencies to be jointly analysed (Isaac et al., 2020), relying on mathematical links between the distributions used to model each currency.

Novel survey methods may add additional complications when considering ecological currencies. For example, DNA sequences from molecular surveys usually do not quantify abundance. In addition, detection probabilities from molecular data can be variable (Buxton et al., 2018), complicating assessments of non-detections. As yet, models for integrating molecular data with traditional sampling data are still in development but are a promising future direction for freshwater analyses.

2.3.2 | Sampling and sample processing protocols

The sampling protocol should define all of the methods, techniques, approaches and equipment that are used to generate observations, as well as the quality-assurance steps in place. Differences among protocols need to be considered when integrating data. For example, kick-sampling and dredge net survey protocols will deliver different, but related, information on species richness, species abundances and community composition of macroinvertebrates when conducted in the same river at the same time, because each method samples the habitat differently (Moore & Murphy, 2015).

A particular challenge for data integration occurs when sampling protocols differ in their spatial or temporal units. For example, a macrophyte point sample and a transect might deliver useful information on the same stretch of river or lake, but the area covered by the latter is far greater. Similar problems occur when integrating data collected at different times of year, or at different depths in the water column. In statistics, this is known as the “change of support” problem and requires additional modelling steps to be taken when constructing integrated models (Pacifi et al., 2019).

2.3.3 | Sampling design

The sampling design of any monitoring activity covers the rules underpinning where and when samples are taken. Sampling designs for monitoring schemes fall along a gradient, from those determined mainly by scientific rationale to those that are driven by practical and/or logistical considerations. Ideal sampling would be representative across the spatial domain of interest and at a consistent temporal frequency over the long-term; one which is suited to disentangling the decadal, interannual, seasonal and

short-term timescales over which ecological dynamics manifest (Amundsen et al., 2019; Perga et al., 2018; Ryo et al., 2019; Seebens et al., 2007).

There are, however, many scenarios whereby practical or logistical constraints dictate changes to pre-planned, structured surveys. Within biodiversity monitoring, this can often be the case with individual species recording, particularly rare species, where a fully random location selection routine could return very few occurrence records. In such circumstances opportunistic data, with no underlying design, may provide a far richer data source with more information content.

Sampling design is important to consider in integrated modelling if one or more datasets are collected in a way that is not random, particularly if there is any evidence that data might be biased towards certain areas or time periods. To understand any biases a "Risk of Bias" assessment can help to identify whether spatial, temporal or other biases arising through surveyor choice could affect model results (Boyd et al., 2022). Failing to account for these biases in integrated modelling can lead to models that are less useful than single dataset models (Simmonds et al., 2020).

2.4 | Integrated analysis: An example with *Agabus* water beetles

In order to illustrate how an integrated analysis can be achieved using the datasets described in Table 1 we create a model of *Agabus* beetle distributions across Great Britain. This example is intended to demonstrate the methodological aspects of integrating datasets, the

mechanics of doing so and the potential impact, and is not intended as a comprehensive assessment of the distribution of *Agabus* beetles. *Agabus* was chosen to illustrate an integrative analysis as a result of the large number of records and relative consistency in nomenclature between datasets. We show that the integrated analysis benefits from the robust, structured design of the BIOSYS data, which offers unbiased, representative sampling across the whole domain, and the large sample size of occurrence records available from the citizen science WBS data (Figure 2).

The distribution of data pooled over a 3 year period (2001–2003, chosen as it represents the most recent period for which data are publicly available across both schemes) clearly illustrates the different spatial extent offered by each data source (Figure 2a,b). BIOSYS data cover England only whilst the WBS data cover the whole of Great Britain. If the mismatch in spatial extent was ignored within the integrated analysis, then undue weight could be given to the WBS data over the BIOSYS data purely because there are large regions where these are the only data available. To mitigate this, we restricted the spatial extent of the analysis to England, where we had records from both datasets.

The datasets also recorded different ecological currencies (counts vs. presence-only). For this case study our aim was to model occurrence, so we simplified the count data to presence–absence. We accounted for the protocol differences between the schemes by fitting a dataset-specific intercept so that each dataset has a different baseline probability of an *Agabus* beetle being observed. This approach can be used to account for simple differences between datasets (e.g., a difference in water volume captured) or the compound effect of multiple differences, as with the water beetle data.

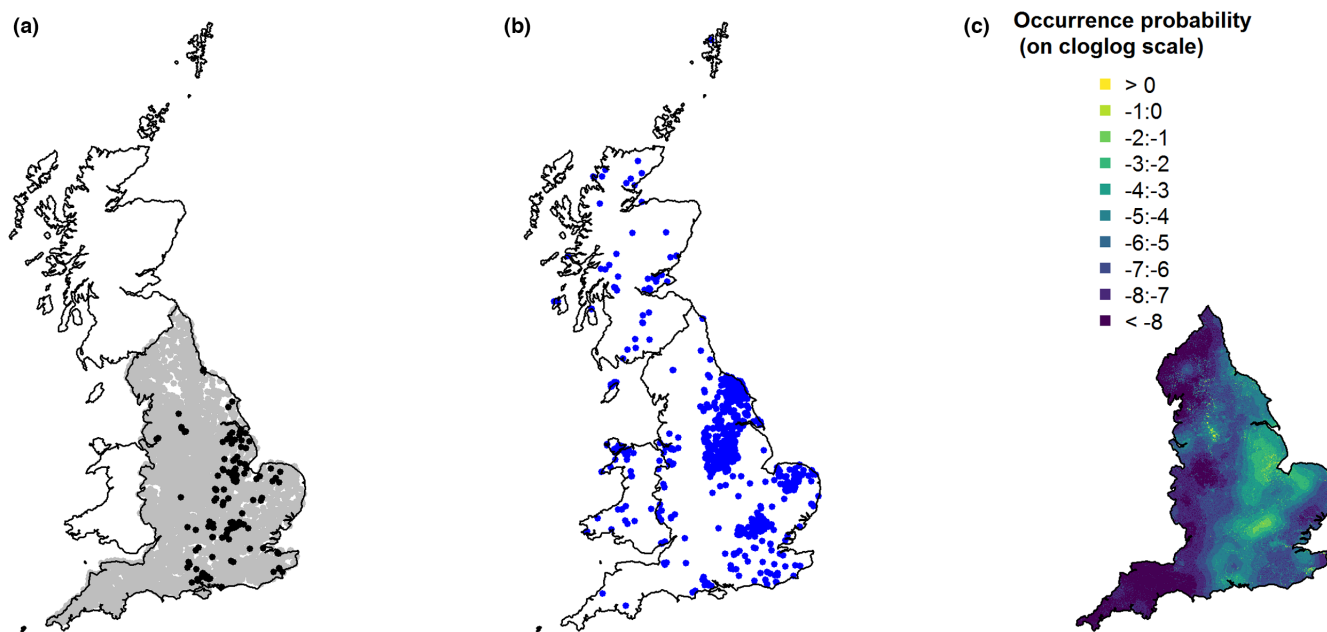


FIGURE 2 The spatial distribution of *Agabus* water beetles from (a) EA BIOSYS, with absences shown in grey and presences in black (157 presences), and (b) Water Beetle Surveys (1086 observations) between 2001 and 2003, and (c) predicted distribution of *Agabus* water beetles at 1-km resolution after integrated modelling using both datasets.

We included four demonstrative environmental covariates in our model relating to the number of nodes and sources on the river network and urbanisation (Appendix S1). Although these covariates were chosen as a consequence of their potential impact on water beetles, we acknowledge there are many other potential drivers and we did not conduct an exhaustive review of covariates for this exemplar. We assumed that the effects of environmental covariates would be the same across both datasets. We also assumed that there was a shared spatial pattern in the data – both datasets would reflect that some parts of the country contained waterways more likely to be occupied by *Agabus* beetles. The WBS data are clustered into areas of intensive observation and areas with little or no sampling effort (Figure 2b), a pattern not shared with the BIOSYS data. We therefore included an additional spatial term to account for uneven sampling effort in the WBS data.

The model structure (Figure 3) demonstrates which aspects of the model are assumed to be shared between datasets (environmental covariate effects and shared spatial process) and which are allowed to vary between datasets (mean and error, WBS-only spatial term). A detailed description of the modelling process and covariates used is available in the Appendix S1. The prediction from the integrated model highlights the areas where *Agabus* beetles are predicted to be most likely to occur based on information in both datasets (Figure 2c). A comparison with a model of BIOSYS data only is presented in Appendix S1, along with uncertainty maps. Integration can increase the power to detect potential driver effects and we found that significant covariate effects on water beetle occurrence were identified using the integrated model that could not be identified using only BIOSYS data (Table 2; Appendix S1). The SD associated with each of the estimated parameters also is smaller in the integrated model than in the BIOSYS data only model, demonstrating that we can estimate

covariate effects with increased confidence in the integrated model (Table 2).

This exemplar demonstrates one approach to integrating these two datasets but note that others are available; for example, we could have included abundance data from BIOSYS directly instead of simplifying to presence–absence (Farr et al., 2021; Zipkin et al., 2017). Although we decided to include only data from England, integrated approaches also can be used where there is spatial mismatch among datasets (Bowler et al., 2019). Our exemplar estimates patterns over space, but we could equally consider an integrated trend analysis over time. We also acknowledge that although integrated approaches can allow us to efficiently use available observational data, they may not overcome the general limitations common to any observational studies (e.g., an inability to confidently quantify ecological processes or identify causal relationships) where an experimental approach may be needed. For each application it will be important to carefully consider the goal of the modelling, the dataset properties, and the assumptions made about underlying ecological and observation processes.

3 | CONCLUSIONS

We believe that integrating different sources of freshwater biodiversity data within a common analytical framework presents opportunities for improved understanding of freshwater ecosystems. Integrated models have the potential to provide more robust measures of biodiversity, more power to identify potential underlying drivers at different scales, and an ability to identify and quantify key uncertainties. Integrated approaches also allow us to make fuller use of the increasing and hugely diverse data now available on freshwater systems, by defining a common framework in which

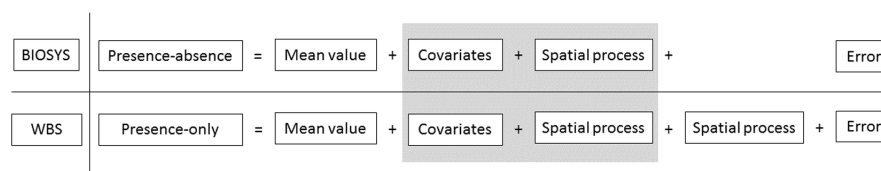


FIGURE 3 Schematic representation of the integrated model fitted, showing the components unique to each scheme and those considered to be common across schemes (those within the grey box). Note that BIOSYS data were downgraded from counts to presence–absence for analysis.

TABLE 2 Estimated coefficients and associated uncertainty for the fixed effects estimated from the integrated model and BIOSYS data only model.

	Integrated model		BIOSYS data only model	
	Estimated coefficient	SD	Estimated coefficient	SD
Density of sources	0.874	0.101	0.409	0.165
Density of nodes	0.087	0.025	−0.072	0.044
Sources:node ratio	−1.860	0.242	−1.028	0.459
Proportion of urban area	0.033	0.006	−0.025	0.018

Note: Entries in bold indicate significant effects for which 95% credible intervals do not overlap zero.

Box 1 Outstanding research questions

- What model structures are needed to integrate novel data sources such as eDNA, image analysis and bioacoustics with more traditional data sources?
- How do we combine data sources to model processes potentially affecting freshwater biodiversity at different spatial and temporal scales?
- What is the best way to evaluate integrated models using multiple datasets?
- How can the complex spatial structure of river and pond networks be accounted for in integrated analyses?

disparate data can be combined. There are many outstanding research questions to be addressed (some highlighted in Box 1) and we hope that this article will serve to catalyse data integration within the freshwater research community, to fully realise the potential of diverse data to enhance fundamental understanding and guide ecosystem restoration.

AUTHOR CONTRIBUTION

Conceptualisation: MF, PH, NI. Developing methods, data analysis: PH, SJ. Preparation of figures and tables: HR, SJ, PH. Data interpretation, conducting the research, writing: MF, PH, NI, SJ, HR, ST, EM, HF.

ACKNOWLEDGEMENTS

This work was funded through a research partnership agreement between Defra and the UK Centre for Ecology & Hydrology (UKCEH) no. 32156, which builds upon work supported by the Natural Environment Research Council award np. NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability.

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

Both of the datasets used in the case study are publicly accessible and usable based on an open government licence and creative commons licence respectively. The EA BIOSYS data can be downloaded from the Ecology and Fish Data Explorer (<https://environment.data.gov.uk/ecology/explorer/>) developed by the EA and using the Defra Data Services Platform infrastructure (<https://environment.data.gov.uk/dataset/fa98090d-d715-4d34-80f9-bb7621aa7101>). The citizen science scheme occurrence records were found, accessed and downloaded as a single text file from the Global Biodiversity Information Facility (GBIF) via the NBN Atlas (<https://registry.nbnatlas.org/public/show/dr686>). Code to conduct the analysis is available at https://github.com/NERC-CEH/EA_Freshwater_Integration.

ORCID

Susan G. Jarvis  <https://orcid.org/0000-0001-5382-5135>

Hannah A. Risser  <https://orcid.org/0000-0001-9819-1092>

REFERENCES

- Amundsen, P. A., Primicerio, R., Smalås, A., Henriksen, E. H., Knudsen, R., Kristoffersen, R., & Klemetsen, A. (2019). Long-term ecological studies in northern lakes—Challenges, experiences, and accomplishments. *Limnology & Oceanography*, *64*, S11–S21.
- Balfour-Browne Club. (2020). Water beetle surveys from Britain and Ireland. *Occurrence Dataset*. <https://doi.org/10.15468/npcgrpaccsessedviaGBIF.orgon2021-01-18>
- Bishop, I. J., Head, J., Shepherd, C., Hayes, S., Loisselle, S., Scott-Somme, K., Stevenson, C., van Noordwijk, T., Watson, A., Fitch, B., Brooks, S., Brierley, B., & Fry, M. (2021). *The role of citizen science in UK freshwater monitoring*. Earthwatch Europe.
- Bowler, D. E., Nilsen, E. B., Bischof, R., O'Hara, R. B., Yu, T. T., Oo, T., Aung, M., & Linnell, J. D. C. (2019). Integrating data from different survey types for population monitoring of an endangered species: The case of the Eld's deer. *Scientific Reports*, *9*, 7766. <https://doi.org/10.1038/s41598-019-44075-9>
- Boyd, R. J., Powney, G. D., Burns, F., Danet, A., Duchenne, F., Grainger, M. J., Jarvis, S. G., Martin, G., Nilsen, E. B., Porcher, E., Stewart, G. B., Wilson, O. J., & Pescott, O. L. (2022). ROBITT: A tool for assessing the risk-of-bias in studies of temporal trends in ecology. *Methods in Ecology and Evolution*, *13*, 1497–1507. <https://doi.org/10.1111/2041-210X.13857>
- Brooks, S. J., Fitch, B., Davy-Bowker, J., & Alvarez, C. S. (2019). Anglers' Riverfly monitoring initiative (ARMI): A UK-wide citizen science project for water quality assessment. *Freshwater Science*, *38*, 270–280.
- Buckland, S. T., & Johnston, A. (2017). Monitoring the biodiversity of regions: Key principles and possible pitfalls. *Biological Conservation*, *214*, 23–34.
- Buxton, A. S., Groombridge, J. J., & Griffiths, R. A. (2018). Seasonal variation in environmental DNA detection in sediment and water samples. *PLoS One*, *13*, e0191737. <https://doi.org/10.1371/journal.pone.0191737>
- DeWan, A. A., & Zipkin, E. F. (2010). An integrated sampling and analysis approach for improved biodiversity monitoring. *Environmental Management*, *45*, 1223–1230.
- Elliott, J. A., Henrys, P., Tanguy, M., Cooper, J., & Maberly, S. C. (2015). Predicting the habitat expansion of the invasive roach *Rutilus rutilus* (Actinopterygii, Cyprinidae), in Great Britain. *Hydrobiologia*, *751*, 127–134.
- Farr, M. T., Green, D. S., Holekamp, K. E., & Zipkin, E. F. (2021). Integrating distance sampling and presence-only data to estimate species abundance. *Ecology*, *102*, e03204. <https://doi.org/10.1002/ecy.3204>
- Fletcher, R. J., Jr., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, *100*, e02710.
- Foster, G. N. (2015). Taking the oldest insect recording scheme into the 21st century. *Biological Journal of the Linnean Society*, *115*, 494–504.
- Harper, L. R., Buxton, A. S., Rees, H. C., Bruce, K., Brys, R., Halfmaerten, D., Read, D. S., Watson, H. V., Sayer, C. D., Jones, E. P., Priestley, V., Mächler, E., Múrria, C., Garcés-Pastor, S., Medupin, C., Burgess, K., Benson, G., Boonham, N., Griffiths, R. A., ... Hänfling, B. (2019). Prospects and challenges of environmental DNA (eDNA) monitoring in freshwater ponds. *Hydrobiologia*, *826*, 25–41.
- Henrys, P. A., & Jarvis, S. G. (2019). Integration of ground survey and remote sensing derived data: Producing robust indicators of habitat

- extent and condition. *Ecology and Evolution*, 9, 8104–8112. <https://doi.org/10.1002/ece3.5376>
- Huang, J., & Frimpong, E. A. (2016). Limited transferability of stream-fish distribution models among river catchments: Reasons and implications. *Freshwater Biology*, 61, 729–744.
- Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, 35, 56–67.
- Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, 5, 1052–1060. <https://doi.org/10.1111/2041-210X.12254>
- Jarić, I., Roll, U., Arlinghaus, R., Belmaker, J., Chen, Y., China, V., Douda, K., Essl, F., Jähnig, S. C., Jeschke, J. M., Kalinkat, G., Kalous, L., Ladle, R., Lennox, R. J., Rosa, R., Sbragaglia, V., Sherren, K., Šmejkal, M., Soriano-Redondo, A., ... Correia, R. A. (2020). Expanding conservation culturomics and iEcology from terrestrial to aquatic realms. *PLoS Biology*, 18(10), e3000935. <https://doi.org/10.1371/journal.pbio.3000935>
- Jenny, J.-P., Anneville, O., Arnaud, F., Baulaza, Y., Bouffard, D., Domaizon, I., Bocaniov, S. A., Chèvre, N., Ditttrich, M., Dorioz, J.-M., Dunlop, E. S., Dur, G., Guillard, J., Guinaldo, T., Jacquet, S., Jamoneau, A., Jawed, Z., Jeppesen, E., Krantzberg, G., ... Weyhenmeyer, G. A. (2020). Scientists' warning to humanity: Rapid degradation of the world's large lakes. *Journal of Great Lakes Research*, 46, 686–702.
- Kakouei, K., Kraemer, B. M., Anneville, O., Carvalho, L., Feuchtmayr, H., Graham, J. L., Higgins, S., Pomati, F., Rudstam, L. G., Stockwell, J. D., Thackeray, S. J., Vanni, M. J., & Adrian, R. (2021). Phytoplankton and cyanobacteria abundances in mid-21st century lakes depend strongly on future land use and climate projections. *Global Change Biology*, 27, 6409–6422. <https://doi.org/10.1111/gcb.15866>
- Linke, S., Gifford, T., Desjonquères, C., Tonolla, D., Aubin, T., Barclay, L., Karaconstantis, C., Kennard, M. J., Rybak, F., & Sœur, J. (2018). Freshwater ecoacoustics as a tool for continuous ecosystem monitoring. *Frontiers in Ecology and the Environment*, 16, 231–238.
- Martino, S., Pace, D. S., Moro, S., Casoli, E., Ventura, D., Frachea, A., Silvestri, M., Arcangeli, A., Giacomini, G., Ardizzone, G., & Jona Lasinio, G. (2021). Integration of presence-only data from several sources: A case study on dolphins' spatial distribution. *Ecography*, 44, 1533–1543. <https://doi.org/10.1111/ecog.05843>
- Metcalfe, A. N., Kennedy, T. A., Mendez, G. A., & Muehlbauer, J. D. (2022). Applied citizen science in freshwater research. *WIREs Water*, 9, e1578. <https://doi.org/10.1002/wat2.1578>
- Miller, D. A., 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.
- Mills, J. A., Teplitsky, C., Arroyo, B., Charmantier, A., Becker Peter, H., Birkhead, T. R., Bize, P., Blumstein, D. T., Bonenfant, C., Boutin, S., Bushuev, A., Cam, E., Cockburn, A., Côté, S. D., Coulson, J. C., Daunt, F., Dingemanse, N. J., Doligez, B., Drummond, H., ... Zedrosser, A. (2015). Archiving primary data: Solutions for long-term studies. *Trends in Ecology & Evolution*, 30, 581–589.
- Moore, I. E., & Murphy, K. J. (2015). Evaluation of alternative macroinvertebrate sampling techniques for use in a new tropical freshwater bioassessment scheme. *Acta Limnologica Brasiliensia*, 27, 213–222.
- Moriarty, M., Sethi, S. A., Pedreschi, D., Smeltz, T. S., McGonigle, C., Harris, B. P., Wolf, N., & Greenstreet, S. P. (2020). Combining fisheries surveys to inform marine species distribution modelling. *ICES Journal of Marine Science*, 77, 539–552.
- Murphy, J., & Weatherby, A. (2008). *Countryside survey technical report No5/07: Freshwater manual v1.0*. Centre for Ecology and Hydrology.
- Nowak, M. M., Dziób, K., Ludwisiak, Ł., & Chmiel, J. (2020). Mobile GIS applications for environmental field surveys: A state of the art. *Global Ecology and Conservation*, 23, e01089.
- Pacifici, K., Reich, B. J., Miller, D. A. W., & Pease, B. S. (2019). Resolving misaligned spatial data with integrated species distribution models. *Ecology*, 100, e02709. <https://doi.org/10.1002/ecy.2709>
- Palmer, S. C. J., Kutser, T., & Hunter, P. D. (2015). Remote sensing of inland waters: Challenges, progress and future directions. *Remote Sensing of Environment*, 157, 1–8.
- Pellerin, B. A., Stauffer, B. A., Young, D. A., Sullivan, D. J., Bricker, S. B., Walbridge, M. R., Clyde, G. A., & Shaw, D. M. (2016). Emerging tools for continuous nutrient monitoring networks: Sensors advancing science and water resources protection. *Journal of the American Water Resources Association*, 52, 993–1008.
- Perga, M.-E., Bruel, R., Rodriguez, L., Guénand, Y., & Bouffard, D. (2018). Storm impacts on alpine lakes: Antecedent weather conditions matter more than the event intensity. *Global Change Biology*, 24, 5004–5016.
- Piepho, H. P., & Ogotu, J. O. (2002). A simple mixed model for trend analysis in wildlife populations. *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 350–360.
- Read, D. S., Bowes, M. J., Newbold, L. K., & Whiteley, A. S. (2014). Weekly flow cytometric analysis of riverine phytoplankton to determine seasonal bloom dynamics. *Environmental Science: Processes & Impacts*, 16, 594–603.
- Reaney, S. M., Mackay, E. B., Haygarth, P. M., Fisher, M., Molineux, A., Potts, M., & Benskin, C. M. H. (2019). Identifying critical source areas using multiple methods for effective diffuse pollution mitigation. *Journal of Environmental Management*, 250, 109366. <https://doi.org/10.1016/j.jenvman.2019.109366>
- Reichman, O. J., Jones, M., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, 331, 703–705. <https://doi.org/10.1126/science.1197962>
- Revenga, C., Campbell, I., Abell, R., de Villiers, P., & Bryer, M. (2005). Prospects for monitoring freshwater ecosystems towards the 2010 targets. *Philosophical Transactions of the Royal Society B*, 360, 397–413.
- Rusak, J. A., Tanentzap, A. J., Klug, J. L., Rose, K. C., Hendricks, S. P., Jennings, E., Laas, A., Pierson, D., Ryder, E., Smyth, R. L., White, D. S., Winslow, L. A., Adrian, R., Arvola, L., de Eyto, E., Feuchtmayr, H., Honti, M., Istvánovics, V., Jones, I. D., ... Zhu, G. (2018). Wind and trophic status explain within and among-lake variability of algal biomass. *Limnology and Oceanography Letters*, 3, 409–418. <https://doi.org/10.1002/lol2.10093>
- Ryo, M., Aguilar-Trigueros, C. A., Pinek, L., Muller, L. A. H., & Rillig, M. C. (2019). Basic principles of temporal dynamics. *Trends in Ecology & Evolution*, 34, 723–733.
- Schmidt-Kloiber, A., Bremerich, V., De Wever, A., Jähnig, S. C., Martens, K., Strackbein, J., Tockner, K., & Hering, D. (2019). The freshwater information platform: A global online network providing data, tools and resources for science and policy support. *Hydrobiologia*, 838, 1–11. <https://doi.org/10.1007/s10750-019-03985-5>
- Seebens, H., Straile, D., Hoegg, R., Stich, H.-B., & Einsle, U. (2007). Population dynamics of a freshwater calanoid copepod: Complex responses to changes in trophic status and climate variability. *Limnology & Oceanography*, 52, 2364–2372.
- Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., & O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, 43, 1413–1422.
- Thackeray, S. J., & Hampton, S. E. (2020). The case for research integration, from genomics to remote sensing, to understand biodiversity change and functional dynamics in the world's lakes. *Global Change Biology*, 26, 3230–3240.
- Tickner, D., Opperman, J. J., Abell, R., Acreman, M., Arthington, A. H., Bunn, S., Cooke, S. J., Dalton, J., Darwall, W., Edwards, G.,

- Harrison, I., Hughes, K., Jones, T., Leclère, D., Lynch, A. J., Leonard, P., McClain, M. E., Muruven, D., Olden, J. D., ... Young, L. (2020). Bending the curve of global freshwater biodiversity loss: An emergency recovery plan. *Bioscience*, 70, 330–342.
- Tulloch, A. I. T., Auerbach, N., Avery-Gomm, S., Bayraktarov, E., Butt, N., Dickman, C. R., Ehmke, G., Fisher, D. O., Grantham, H., Holden, M. H., Lavery, T. H., Leseberg, N. P., Nicholls, M., O'Connor, J., Roberson, L., Smyth, A. K., Stone, Z., Tulloch, V., Turak, E., ... Watson, J. E. M. (2018). A decision tree for assessing the risks and benefits of publishing biodiversity data. *Nature Ecology and Evolution*, 2, 1209–1217. <https://doi.org/10.1038/s41559-018-0608-1>
- Turner, R. K., Griffiths, R. A., Wilkinson, J. W., Julian, A. M., Toms, M. P., & Isaac, N. J. B. (2023). Diversity, fragmentation, and connectivity across the UK amphibian and reptile data management landscape. *Biodiversity and Conservation*, 32, 37–64. <https://doi.org/10.1007/s10531-022-02502-w>
- van Rees, C. B., Waylen, K. A., Schmidt-Kloiber, A., Thackeray, S. J., Kalinkat, G., Martens, K., Domisch, S., Lillebø, A. I., Hermoso, V., Grossart, H.-P., Schinegger, R., Decler, K., Adriaens, T., Denys, L., Jarić, I., Janse, J. H., Monaghan, M. T., De Wever, A., Geijzendorffer, I., ... Jähnig, S. C. (2021). Safeguarding freshwater life beyond 2020: Recommendations for the new global biodiversity framework from the European experience. *Conservation Letters*, 14, e12771. <https://doi.org/10.1111/conl.12771>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., & Bouwman, J. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 1–9.
- Zipkin, E. F., Rossman, S., Yackulic, C. B., Wiens, J. D., Thorson, J. T., Davis, R. J., & Grant, E. H. C. (2017). Integrating count and detection–nondetection data to model population dynamics. *Ecology*, 98, 1640–1650. <https://doi.org/10.1002/ecy.1831>
- Zipkin, E. F., Zylstra, E. R., Wright, A. D., Saunders, S. P., Finley, A. O., Dietze, M. C., Itter, M. S., & Tingley, M. W. (2021). Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, 19, 30–38.
- Zulian, V., Miller, D. A. W., & Ferraz, G. (2021). Integrating citizen-science and planned-survey data improves species distribution estimates. *Diversity and Distributions*, 27, 2498–2509. <https://doi.org/10.1111/ddi.13416>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jarvis, S. G., Mackay, E. B., Risser, H. A., Feuchtmayr, H., Fry, M., Isaac, N. J. B., Thackeray, S. J., & Henrys, P. A. (2023). Integrating freshwater biodiversity data sources: Key challenges and opportunities. *Freshwater Biology*, 68, 1479–1488. <https://doi.org/10.1111/fwb.14143>