scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Range maps and waterbody occupancy data for 1158 freshwater macroinvertebrate genera in the contiguous USA

Ethan A. Brown[1]✉, Ronald A. Hellenthal[1], Michael B. Mahon[2], Samantha L. Rumschlag[2] & Jason R. Rohr[1]

Range maps are used to estimate the geographic extent of taxa, providing valuable information for biodiversity and conservation research and management. Freshwater macroinvertebrates are not well-represented in the range map literature relative to freshwater vertebrates. To address this knowledge gap, we provide range maps for 1158 freshwater macroinvertebrate genera based on two decades of publicly available occurrence data from the USEPA National Aquatic Resource Surveys, which included 11,628 sites and 6,906,990 organisms across the contiguous USA. Maps were created by applying unweighted and weighted pair group method with arithmetic mean clustering and single-linkage clustering algorithms to the occurrence data and creating three layers of polygons from the minimum convex hulls of clusters. A total of 25 freshwater macroinvertebrate classes are represented in the range map dataset. Most mapped genera were insects (394/1158), followed by malacostracans (242/1158), polychaetes (182/1158), and bivalves (121/1158). Additionally, we provide waterbody type percent occupancy data for all genera, detailing how genera are partitioned between boatable streams, wadeable streams, inland lakes, Laurentian Great Lakes, and coastal estuaries.

Background & Summary

Taxonomic range maps are commonly used in ecology for spatial modelling[1,2], characterizing species richness[2–4], and aiding in conservation efforts[2,5]. The International Union for Conservation of Nature (IUCN) hosts the largest database of expert-drawn range maps and makes them freely available for research applications[4,6]. This IUCN repository contains thousands of range maps for mammals, birds, reptiles, amphibians, and fishes[6].

Historically, range maps consist of polygons that are hand-drawn by experts and are often based on textual range descriptions, country-level occurrence information, museum records, or a combination of regional-scale maps or recorded occurrences compiled from multiple sources. Using these approaches, maps can be slow and costly to produce and may not be directly derived from standardized occurrence data[2,7,8]. While extensive, the IUCN range map database does not contain maps for certain taxonomic groups, including freshwater benthic macroinvertebrates. Thus, to fill the large gaps in range map availability for underrepresented taxa, it would be useful to automate range map creation by using computational methods in conjunction with occurrence data collected via a spatially balanced, randomized sampling design[1,7,9].

Among the least documented groups in the range map literature are aquatic macroinvertebrates[6]. To address this knowledge gap, we harnessed two decades of macroinvertebrate occurrence data from the USEPA National Aquatic Resource Surveys (NARS)[10] to generate range maps for freshwater macroinvertebrate genera in the contiguous United States. The NARS database includes benthic macroinvertebrate sampling data from (1) the National Rivers and Streams Assessment (NRSA), comprised of three nationwide survey cycles of boatable and wadeable streams occurring from 2008–2009, 2013–2014, and 2018–2019; (2) the National Coastal Condition Assessment (NCCA), comprised of four nationwide survey cycles from both US estuaries and the Great Lakes occurring in 1999–2000, 2005–2006, 2010, and 2015; and (3) the National Lakes Assessment (NLA), comprised of three

[1]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA. [2]U.S. EPA Office of Research and Development, Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division, 6201 Congdon Blvd., Duluth, MN, USA. ✉e-mail: ebrown23@nd.edu

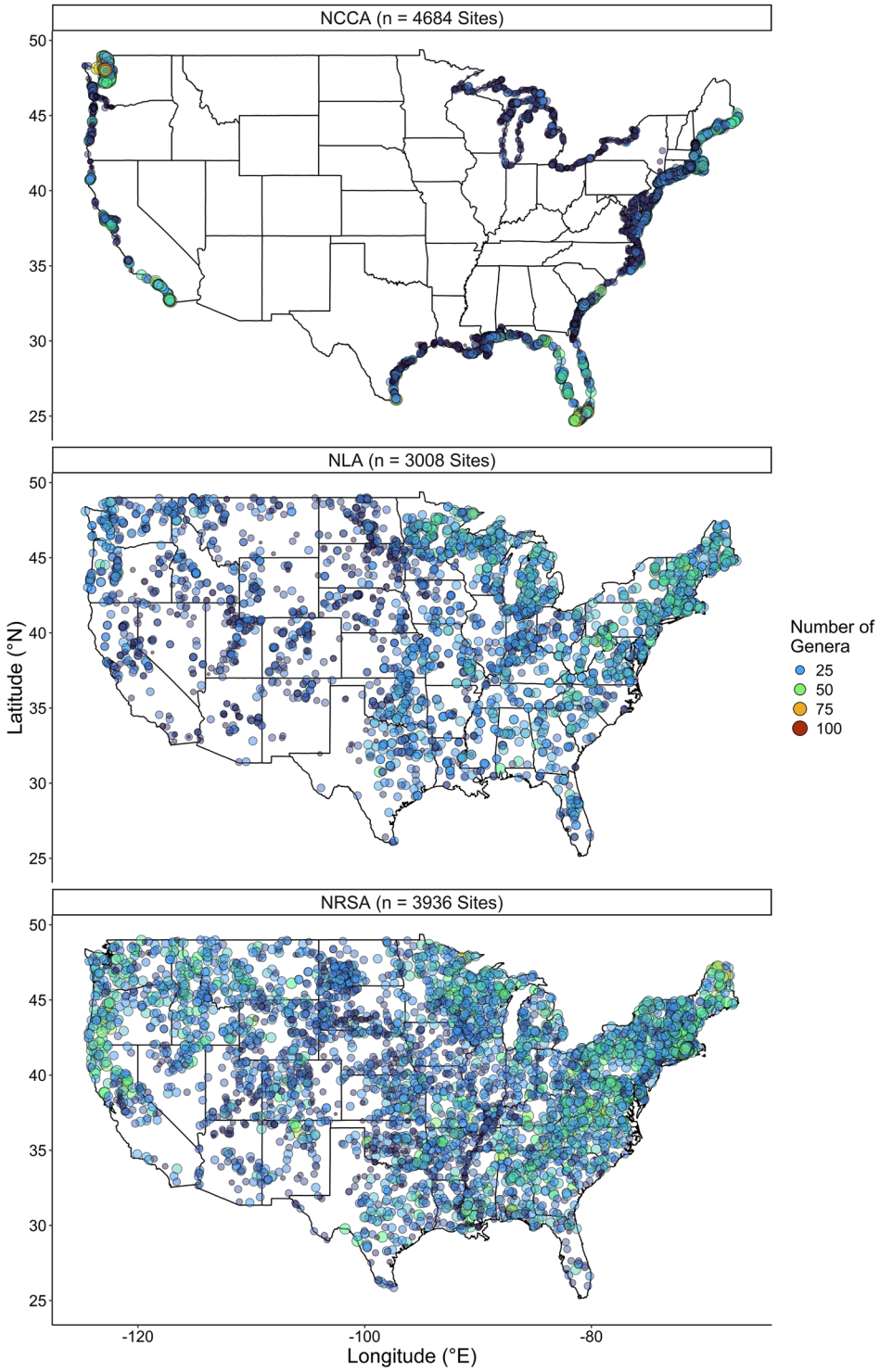

Fig. 1 Number of macroinvertebrate genera detected at each National Aquatic Resource Surveys (NARS) sampling site across all sampling dates. We used all benthic macroinvertebrate data from the National Coastal Condition Assessment (NCCA) (1999–2000, 2005–2006, 2010, 2015), National Lakes Assessment (NLA) (2007, 2012, 2017), and National Rivers and Streams Assessment (NRSA) (2008–2009, 2013–2014, 2018–2019) programs, resulting in a total of 11,628 surveyed sites.

nationwide survey cycles of freshwater lakes occurring in 2007, 2012, and 2017. Each of these NARS programs employed a probabilistic sampling design to provide spatially unbiased data across the contiguous USA[11–13] (Fig. 1).

Range maps consist of three layers of convex polygons, each corresponding to a different spatial scale (broad-scale, region-scale, hotspot). By generating three polygon layers, these maps address a wide range of uses in aquatic research and management by providing users with multiple abstractions for the range of each macroinvertebrate genus.

Methods

Data acquisition. We acquired NLA and NCCA source data through the NARS Data Download Tool, available at https://owshiny.epa.gov/nars-data-download/. We acquired the NRSA data through the 'finsyncR' R package[14] which streamlines the data gathering and cleaning process for the NRSA dataset based on the user's specific needs. The NLA and NCCA data were cleaned using the same methodology as 'finsyncR' (see '*Data Cleaning and Filtering*'). We used all data for benthic macroinvertebrates for each survey cycle of NRSA, NLA, and NCCA.

Data cleaning and filtering. After verifying that the taxonomy in the NARS data were current and accurate (see 'Technical Validation'), we applied filtering criteria to ensure that we were accurately estimating the geographic extent of each genus while also avoiding over- or under-representing the range of any one genus. The goal of filtering was to ensure that range estimates were not biased in favour of certain genera or groups of genera, thus ensuring that the maps are statistically robust and appropriate for use in spatial analysis.

First, to account for differences in sampling effort between NARS sampling events and to ensure that range estimates were consistent between genera, we removed data for a handful of samples with extremely low rates of genus-level identification or area sampled (column 'PropID' in the occurrence data). Specifically, we removed samples where less than 5.5% of organisms were identified to genus. This is an important filtering criterion because it avoids biasing the occurrence data in favor of a small number of identified genera when the sample contained many other specimens that were not identified to genus. The area sampled criterion (column 'AreaSampTot_m2' in the occurrence data) differed based on the sampling methods of each NARS program: for NRSA we removed samples with less than 0.74 m^2 sampled; for NLA we removed samples with less than 2.74 m^2 sampled; and for NCCA we removed samples with less than 0.04 m^2 sampled. These combined criteria removed data for 724 NARS sampling events, or 4.9% of all samples.

Next, because it can be unreliable to generate range maps based on small amounts of occurrence data, we removed data for exceedingly rare genera that had not been detected at least 10 occasions across at least five sampling locations, resulting in the removal of 602 genera from the dataset. Then, we removed data for any taxon that had not been identified to the genus level in the NARS source data or had been ambiguously identified using 'slash' names (i.e. Genus A/Genus B). After that, we identified clusters of genera that have undergone genus lumping or splitting over the survey period (1999–2019) using the methods described in the 'finsyncR' R package documentation[14] and Rumschlag *et al.*[15]. After identifying these clusters or 'lumped' genera, we removed any cluster containing more than two genus names. This criterion resulted in the removal of six lumped groups containing a total of 50 genera, thereby reducing uncertainty due to mapping genera that have undergone frequent taxonomic reclassifications. Maps for a total of 33 'lumped' genera are included in the database.

Finally, to ensure maps were not biased against any major taxonomic groups we performed additional filtering to determine whether (1) making genus-level identifications was less common for certain orders and whether (2) the ability of taxonomists to identify genera within some orders has improved over time. To identify such orders, we calculated the proportion of specimens identified to genus within each order and removed data for orders that showed (1) low overall identifications at the genus level or (2) a steep temporal increase in the proportion of organisms identified to genus (Figs. S1–S3). By excluding these orders, an additional 146 genera were removed from the dataset, thereby reducing the likelihood that range maps represent some orders better than others. The resulting dataset contained occurrence data for 1158 freshwater macroinvertebrate genera. Figure 2 and Table 1 summarize the taxonomy of mapped genera by class, order, and family.

Map and shapefile creation. Using the filtered occurrence data, we generated maps and shapefiles representing the geographic range of each genus (Fig. 3), defined as any location where the genus was detected over the survey period (1999–2019). This process consisted of two primary steps: (1) defining clusters of occurrence data for each genus and (2) creating polygons based on each cluster to estimate the spatial extent of each taxon.

First, we selected cluster and polygon creation methods based on precision and accuracy scores, as calculated through cross-validation (see 'Technical Validation'). Using the results of the cross-validation we determined the optimal polygon generation methods to (1) represent broad-scale patterns (e.g. 'West Coast', 'Midwest', 'East Coast'), (2) represent region-scale patterns (e.g. 'Great Lakes Region', 'Pacific Northwest'), and (3) represent hotspots, i.e. areas where there is a high likelihood of detecting a given genus.

Each of the three polygon layers used a different clustering algorithm selected from the cross-validation results. We used Euclidean distance for all clustering operations. For the broad-scale polygons we used single-linkage (SL) clustering (a.k.a. nearest neighbour). For the region-scale polygons we used weighted pair group method with arithmetic mean (WPGMA). For the hotspot-scale polygons we used unweighted pair group method with arithmetic mean (UPGMA) clusters. WPGMA defines clusters by calculating the average distance between all points in two clusters and only joining these clusters if the average distance falls within a defined distance threshold[16]. UPGMA is similar to WPGMA, except a weighting criteria is also applied when averaging the distance between clusters, giving more weight to clusters with more data. SL clustering is a less conservative form of hierarchical clustering that aggregates groups of points based on a friends-of-friends approach; in other words, points will be iteratively added to a cluster if any points in the cluster are within a defined distance threshold of another point. SL clustering often results in larger polygons that can be more elongate in shape, as opposed to polygons generated through a more conservative algorithm such as UPGMA or WPGMA[16].

Next, we calculated minimum convex polygons for each genus, a common practice in the range maps literature[9,17–20]. Minimum convex polygons, also called convex hulls, represent the smallest polygon around a group of points for which no angle exceeds 180 degrees[20].

Then, using the quantitative selection criteria described in 'Technical Validation', we generated three range map polygon layers. To represent hotspot-scale occurrence, we used polygons with a 200-km UPGMA clustering

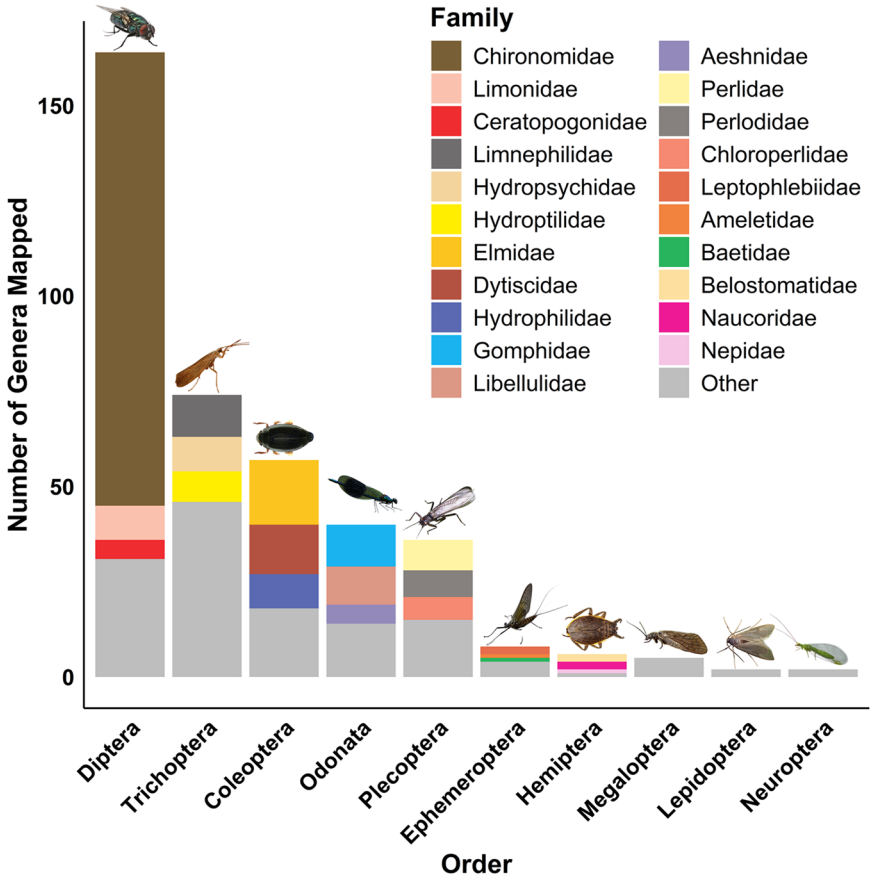

Fig. 2 Summary of insect orders and families represented in the range map database. For each order, the three families with the highest number of genera were labelled with the exception of Megaloptera, Lepidoptera, and Neuroptera which had relatively few genera represented in the National Aquatic Resource Surveys (NARS) data. Insects were the most common taxonomic class mapped as part of this work, representing 394 out of 1158 macroinvertebrate genera. It should be noted that NARS sampling targeted aquatic habitats, thus, genera with terrestrial lifestages will not be fully represented by range maps. Insect images sourced from Wikimedia Commons users Rolf Dietrich Brecher, Udo Schmidt, Dick Belgers, Jakub Halun, Frank Vassen, Andrew Cattoir, Syrio, Ilia Ustyantsev, Biodehio, and Alvesgaspar.

threshold, for region-scale occurrence patterns we used polygons with a 800-km WPGMA clustering threshold; and for broad-scale patterns we used polygons with a 800-km SLC threshold.

Finally, we conducted a two-phase clipping approach on the polygons: phase one involved clipping polygons to the appropriate Hydrologic Unit Code (HUC) boundaries and phase two involved clipping polygons to the relevant NARS sampling region. For phase one, we clipped the regional-scale polygons to include only the basins (6-digit HUC boundaries) where the genus occurred and hotspot-scale polygons to the sub-basins (8-digit HUC boundaries) where the genus occurred. Broad-scale polygons were not clipped to any HUC boundaries. For phase two, we clipped polygons based on which NARS program(s) the occurrence data associated with the polygon originated from. For polygons derived from NLA and/or NRSA data, we clipped the polygons to the border of the contiguous USA, excluding the Great Lakes. For polygons derived from NCCA data only, we clipped polygons to coastal zones of the contiguous USA and the Great Lakes, as defined by the National Oceanic and Atmospheric Administration (NOAA) Coastal Zone Management Act boundaries[21]. For polygons derived from NCCA and either NRSA or NLA data, we clipped polygons to the combined extent of the contiguous US and all coastal or Great Lakes areas.

After generating polygons, we created the final visualization of the range maps by overlaying all three polygon layers for each genus. The maps also depict point-occurrence data and non-detect data (Fig. 3). We provide polygons for each genus as both maps (as *.pdf files) and shapefiles (as *.gpkg files). We also provide occurrence data used to generate maps (as *.csv file).

Waterbody type occupancy data. To supplement the range maps, we calculated the percent waterbody type occupancy for each genus, defined as the relative distribution of a genus among each of the sampled waterbody types as part of the NARS data, adjusted based on sampling effort. The five waterbody types were boatable rivers and streams, wadeable streams, inland lakes, Laurentian Great Lakes, and coastal estuaries. We calculated percent occupancy using Eq. 1,

Class	Number of Orders	Number of Families	Number of Genera
Anthozoa	3	4	4
Ascidiacea	1	2	3
Bivalvia	17	36	121
Branchiopoda	1	1	1
Caudofoveata	1	1	1
Copepoda	2	2	2
Echinoidea	1	3	3
Gastropoda	8	50	92
Hexacorallia	1	4	5
Hexapoda	4	5	9
Holothuroidea	2	2	4
Hoplonemertea	1	6	6
Hydrozoa	3	4	5
Insecta	10	87	394
Malacostraca	7	104	242
Octocorallia	1	3	3
Ophiuroidea	3	5	10
Palaeonemertea	2	3	3
Pilidiophora	1	1	6
Polychaeta	6	36	182
Polyplacophora	1	2	2
Pycnogonida	1	3	3
Scaphopoda	1	2	3
Stenolaemata	1	1	1
Trematoda	1	1	1
incertae sedis	—	20	52

Table 1. Number of orders, families, and genera (maps) within each of the 25 macroinvertebrate classes included in the range map database. Additionally, the database includes 52 genera which are not classified at the Order or Class level, termed '*incertae sedis*.'

$$H_{ij} = \frac{L_{ij}}{L_i} * \frac{L}{L_j} * 100 \quad (1)$$

where H_{ij} is the percent occupancy of genus i in waterbody type j; L_{ij} is the number of NARS sites for waterbody type j where genus i was detected; L_i is the total number of sites where genus i was detected; L is the total number of sites in the NARS dataset; and L_j is the total number of NARS sites for waterbody type j.

After calculating percent occupancy for all genera, we discovered that 62.0% of genera are specialists (taxa that occupy only a single waterbody type); 14.1% of genera occupy two out of five waterbody types; 14.5% of genera occupy three out of five waterbody types; 5.2% of genera occupy four out of five waterbody types; and 4.2% of genera occupy all five waterbody types. Figure 4 shows the distribution and overlap of the number of genera that occupy each waterbody type.

Data Records

All data, code, metadata, shapefiles, and maps are available through the Figshare links referenced below. The repository includes the following items: (1) PDF files containing maps for each macroinvertebrate genus in the NARS dataset, visualized in three ways: a) only range polygons plotted b) range polygons and occurrence data plotted, and c) range polygons, occurrence, and non-detect data plotted[22]; (2) Shapefiles in GeoPackage (*.gpkg) format for all broad-, region-, and hotspot-scale polygons with attribute data describing taxonomic classification of the genus up to the phylum level[23]; (3) cleaned and filtered NARS benthic macroinvertebrate occurrence data used to generate polygons and maps[24]; (4) waterbody type occupancy statistics for each genus[25]; and (5) R code used to create maps and shapefiles[26];, (6) a CSV file listing all genera in the database with taxonomic classification data up to the phylum level[27], and (7) a metadata document describing each of the aforementioned items in greater detail including definitions for all columns and attributes[28].

The occurrence data can be directly linked back to the NARS source data using the columns 'SiteNumber' (called 'SITE_ID' in the source data) and 'CollectionDate' (called 'DATE_COL' in the source data). As part of the data cleaning process, some columns from the source data have been either renamed or removed. All column definitions, in addition to the source code for cleaning NARS data, can be found in Mahon *et al.*[14] and its associated Git repository.

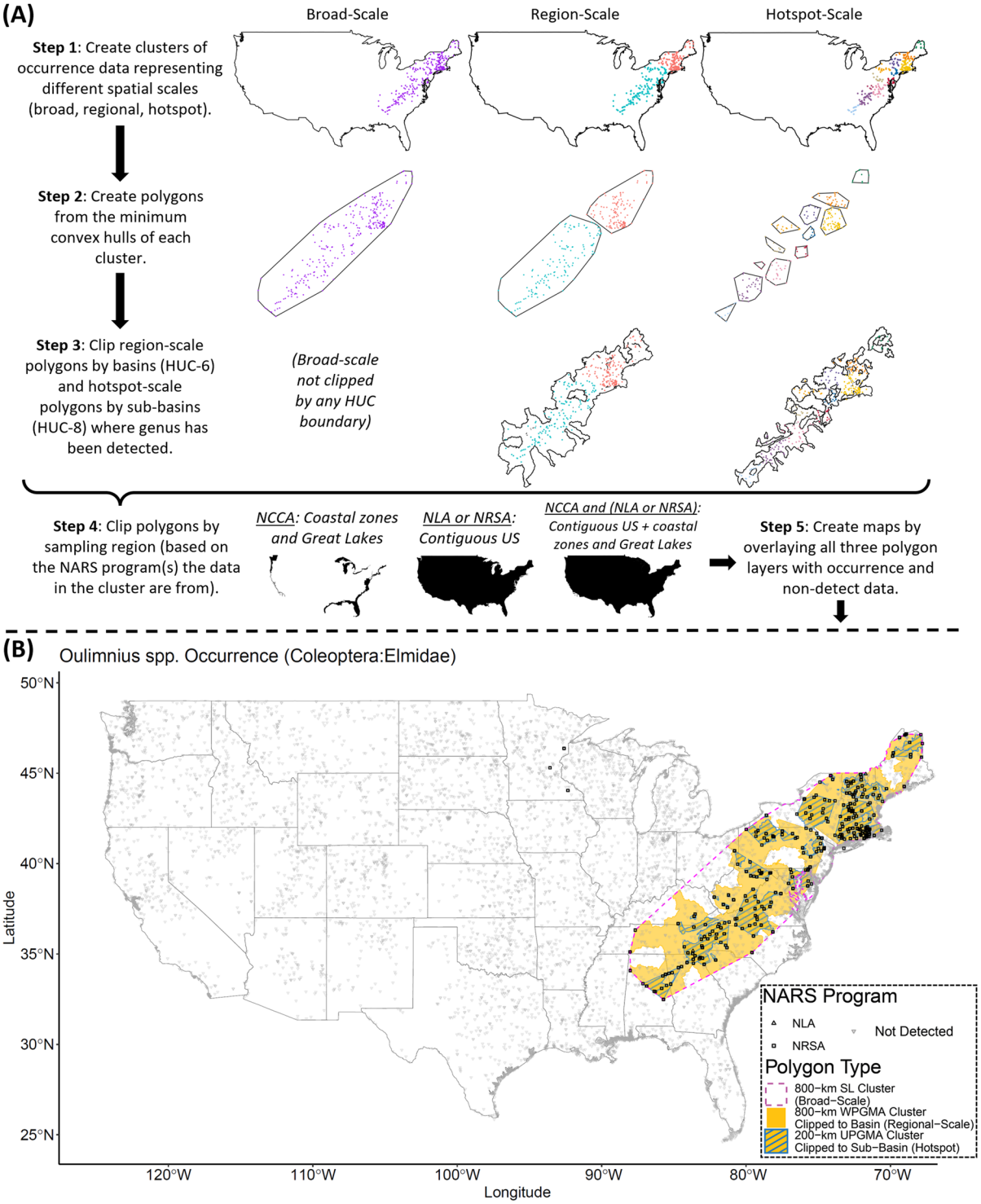

Fig. 3 Range map creation workflow diagram (**A**) and example range map (**B**). The workflow diagram shows each step involved in generating the broad-, regional-, and hotspot-scale polygons, including (1) clustering of occurrence data, (2) creation of polygons around each cluster via the minimum convex hull approach, (3) clipping polygons based on HUC boundaries, and (4) clipping polygons based on sampling region. The maps were generated by plotting all three polygon layers with the occurrence and absence (or non-detect) data. The shape of each point indicates the National Aquatic Resource Surveys (NARS) program associated with the occurrence record. For guidance on use and interpretation of each polygon layer, see 'Interpretation and Use of Each Polygon Layer'. For details on how each layer of polygons was created, see 'Map and Shapefile Creation'.

Technical Validation

Taxonomy QA. To confirm that all genus names were current and correct, we cross checked all genus names in the NARS dataset with both the Global Biodiversity Information Facility (GBIF) database[29], the

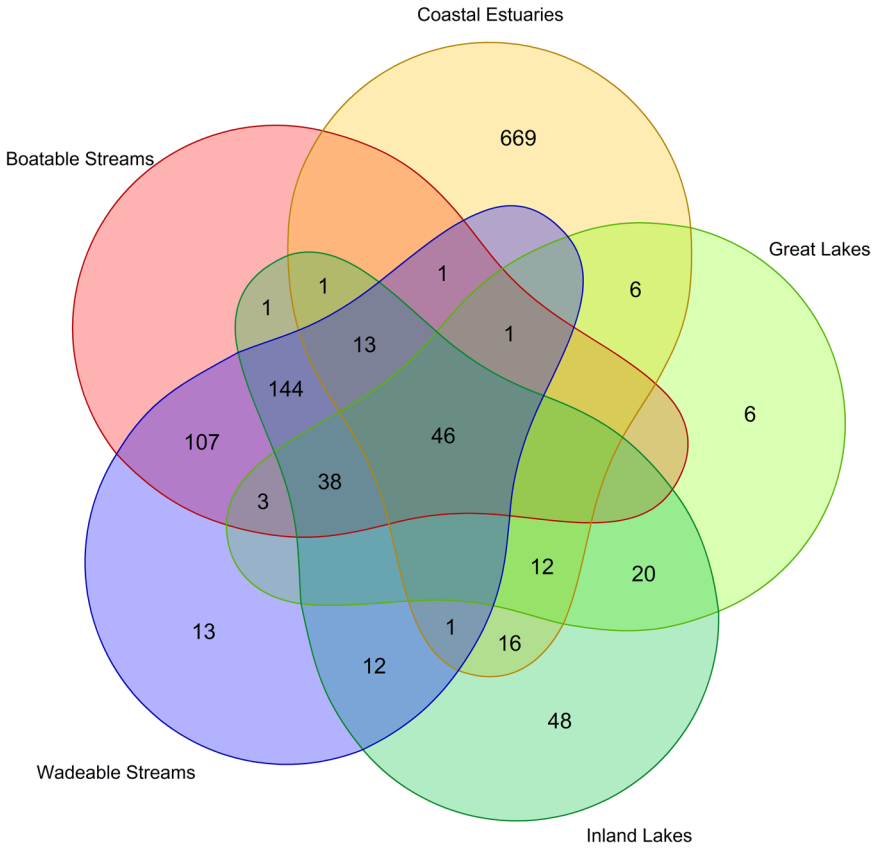

Fig. 4 Venn diagram showing the distribution of genera between the waterbody types represented in the National Aquatic Resource Surveys (NARS) data. Values represent the number of genera that occupy waterbody type or types.

National Center for Biotechnology Information (NCBI) database[30], the Integrated Taxonomic Information System (ITIS) database[31] using the 'taxize' package in R[32]. Additionally, we performed taxonomy checks using MolluscaBase[33] and the World Register of Marine Species (WoRMS)[34], both of which are considered international authorities on up-to-date taxonomic classification information and can also be useful in identifying synonymous genus names. Using these databases, we were able to identify and resolve multiple instances of synonymous genus names, outdated higher classifications of genera, and genus names that are no longer recognized as legitimate. We removed data for any genus that did not return an exact match from any of the aforementioned databases.

Clustering method cross validation. To determine the most appropriate methods for generating polygons in range maps, we randomly separated occurrence data into training (90% of data) and testing (remaining 10% of data) datasets and conducted cross validation. Cross validation consisted of (1) defining clusters of occurrence data based on the training dataset, (2) creating polygons based on those clusters, and (3) calculating accuracy and precision scores for the clustering method using the training polygons and the testing dataset, as described below.

For our purposes we define the *accuracy score* of a clustering method as the proportion of data points in the training data that fell within the polygons generated from the testing data, i.e. Equation 2:

$$Accuracy\ Score = \frac{N_{poly}}{N_{test}} \tag{2}$$

where N_{test} is the size of the test dataset and N_{poly} is the number of records in the test dataset that fell within the training polygons. Thus, accuracy scores represent the reliability with which polygons predict the occurrences of a genus. We define the *precision score* of a clustering method as follows:

$$Precision\ Score = 1 - \frac{E_{poly}}{E_{tot}} \tag{3}$$

where E_{poly} is the number of non-detects that fell within the training polygons, or the number of NARS sites where the genus was not detected, and E_{tot} is the total number of non-detects. The precision score of a clustering method represents the degree to which the polygons exclude sites where the genus has never been detected.

We calculated accuracy and precision scores for a total of 70 unique clustering methods, testing seven clustering algorithms (SL, UPGMA, WPGMA, Ward, CL, UPGMC, WPGMC) crossed with ten distance thresholds (25, 50, 100, 200, 300, 400, 500, 600, 700, 800 km). First, we calculated the accuracy and precision scores for each of the 1158 genera in the NARS data. For each genus, we ran 50 simulations of the cross validation, re-randomizing the training and testing data for each iteration. Then, we averaged the accuracy and precision scores from the 50 simulations to get the genus-level accuracy and precision for the clustering method. Finally, we averaged the genus-level accuracy and precision scores ($n = 1158$) to get the overall accuracy and precision scores for each clustering method. The results of our cross validation are depicted in the supplement of this manuscript (Figs. S4, S5).

For the broad-scale polygons we used the clustering method with the highest accuracy score without a precision score less than 0.25. SL clustering with a threshold of 800 km met these criteria with an accuracy score of 0.74, the highest accuracy of all tested methods, and a precision score of 0.8.

For the region-scale polygons we selected the polygon method with the highest combined accuracy and precision score with an accuracy score greater than 0.5. We excluded SL clustering from consideration for region-scale polygons because SL clustering is a less conservative approach than many of the average-linkage methods and can lead to large, elongate polygons that often do not accurately represent region-scale occurrence patterns. Thus, the clustering method that best met the criteria for region-scale polygons was WPGMA clustering with a threshold of 800 km which had an accuracy score of 0.68 and a precision score of 0.78.

Finally, to select a method for hotspot-scale polygons, we chose the polygon method with the highest precision, without a decrease in accuracy of more than 0.1 from the next highest distance threshold and did not have a precision score below 0.25. For example, if a clustering method had an accuracy of 0.5 at a 300 km threshold, 0.41 at a 200 km threshold, and 0.3 at 100 km, then the 200 km clustering method would be selected, as the 100 km method resulted in an accuracy decrease of 0.11 which violates the criterion. Based on these criteria, the best clustering method for the hotspot-scale was UPGMA clustering with a 200 km threshold. This result aligns with previous work that has demonstrated that the optimal spatial resolution for hotspot delineation is approximately 200 km^3.

Usage Notes

The macroinvertebrate range maps accompanying this manuscript were generated using every location where each genus was detected across the entire temporal extent of the dataset. Users who wish to know the geographic range of a genus within a specific timeframe can generate new maps using the provided R code with the NARS data, filtering occurrence data to include only the timeframe of interest, however, because new sites were selected for each NARS survey cycles, users should be aware that timeframe-specific maps may not estimate genus ranges as extensively or accurately as the time-aggregated maps we provide.

It is also possible to generate maps for higher levels of biological organization. The occurrence data contain taxonomic information for each genus up to the phylum level. To generate these higher-order maps, users should aggregate data based on their desired level of biological organization, then re-generate polygons using the cross-validation and clustering methods described in this manuscript.

Interpretation and use of each polygon layer. Each range map is composed of three polygon layers, each providing an abstraction of the genus range at a different spatial scale. When interpreting polygons or using the associated shapefiles for modelling or statistical analysis, users should be aware of the limitations of each layer. For example, when performing analyses at large spatial scales (e.g. estimating genus richness across the US), hotspot-scale polygons may provide an overly-conservative estimate of genus ranges as compared to region- or broad-scale polygons. Conversely, broad- or regional-scale polygons may not be appropriate for analyses at smaller scales (e.g. predicting macroinvertebrate assemblages at specific locations).

Range map limitations and caveats. When using the provided maps and shapefiles, users should be aware of a few key limitations: First, the NARS data were mostly collected during summer, thus, genera that are more abundant during colder seasons may not be well-represented by these data (e.g., genera within Taeniopterygidae [winter stoneflies]). Additionally, taxa that primarily rely on temporary waters as a breeding habitat, e.g. *Aedes spp*., may not be well-represented if the NARS sampling did not occur during dry periods when these habitats are not present. Finally, because NARS sampling only included aquatic environments, genera with a terrestrial adult lifestage (e.g. many orders of insects including Ephemeroptera, Plecoptera, and Odonata) will not be fully represented by range estimates, in other words, these maps only estimate the geographic extent of immature lifestages.

Code availability

All code for map creation is provided in the Figshare repository[26] that accompanies this manuscript. We generated maps and shapefiles using packages 'sp 1.6-0'[35], 'sf 1.0-9'[36], and 'raster 3.6-14'[37] in R version 4.2.2[38] using RStudio version 2021.9.2.382[39].

Received: 9 April 2024; Accepted: 29 August 2024;
Published online: 12 September 2024

References

1. Jetz, W., Sekercioglu, C. H. & Watson, J. E. Ecological correlates and conservation implications of overestimating species geographic ranges. *Conserv. Biol.* **22**, 110–119, https://doi.org/10.1111/j.1523-1739.2007.00847.x (2008).
2. Marsh, C. J. *et al*. Expert range maps of global mammal distributions harmonised to three taxonomic authorities. *J. Biogeogr.* **49**, 979–992, https://doi.org/10.1111/jbi.14330 (2022).
3. Hurlbert, A. H. & Jetz, W. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *PNAS* **104**, 13384–13389, https://doi.org/10.1073/pnas.0704469104 (2007).
4. Marechaux, I., Rodrigues, A. S. L. & Charpentier, A. The value of coarse species range maps to inform local biodiversity conservation in a global context. *Ecography* **40**, 1166–1176, https://doi.org/10.1111/ecog.02598 (2017).
5. O'Connor, D. *et al*. Updated geographic range maps for giraffe, Giraffa spp., throughout sub-Saharan Africa, and implications of changing distributions for conservation. *Mamm. Rev.* **49**, 285–299, https://doi.org/10.1111/mam.12165 (2019).
6. IUCN. *Spatial Data and Mapping Resources*. *IUCN* https://www.iucnredlist.org/resources/spatial-data-download (2022).
7. Huang, R. M. *et al*. Batch-produced, GIS-informed range maps for birds based on provenanced, crowd-sourced data inform conservation assessments. *Plos One* **16**, https://doi.org/10.1371/journal.pone.0259299 (2021).
8. IUCN. *Mapping Standards and Data Quality for IUCN Red List Spatial Data*. *IUCN* https://www.iucnredlist.org/resources/mappingstandards (2021).
9. Joppa, L. N. *et al*. Impact of alternative metrics on estimates of extent of occurrence for extinction risk assessment. *Conserv. Biol.* **30**, 362–370, https://doi.org/10.1111/cobi.12591 (2016).
10. USEPA Office of Water and Office of Research and Development. Manuals Used in the National Aquatic Resource Surveys. *USEPA* https://www.epa.gov/national-aquatic-resource-surveys/manuals-used-national-aquatic-resource-surveys (2007).
11. USEPA Office of Water and Office of Research and Development. NRSA 2008-09 Field Operations Manual. *USEPA* https://www.epa.gov/national-aquatic-resource-surveys/national-rivers-streams-assessment-2008-2009-field-operations (2008).
12. USEPA Office of Water and Office of Research and Development. NCCA 2015 Field Operations Manual. *USEPA* https://www.epa.gov/national-aquatic-resource-surveys/national-coastal-condition-assessment-2015-field-operations-0 (2015).
13. USEPA Office of Water and Office of Research and Development. NLA 2017 Field Operations Manual. *USEPA* https://www.epa.gov/national-aquatic-resource-surveys/national-lakes-assessment-2017-field-operations-manual (2017).
14. Mahon, M. B. *et al*. finsyncR, an R package to synchronize 27 years of fish and invertebrate data across the United States. *bioRxiv*, https://doi.org/10.1101/2024.02.22.581615 (2024).
15. Rumschlag, S. L. *et al*. Density declines, richness increases, and composition shifts in stream macroinvertebrates. *Sci. Adv.* **9**, eadf4896, https://doi.org/10.1126/sciadv.adf4896 (2023).
16. Kreft, H. & Jetz, W. A framework for delineating biogeographical regions based on species distributions. *J. Biogeogr.* **37**, 2029–2053 (2010).
17. Graham, C. H. & Hijmans, R. J. A comparison of methods for mapping species ranges and species richness. *Glob. Ecol. Biogeogr.* **15**, 578–587 (2006).
18. Hatchwell, B., Anderson, C., Ross, D., Fowlie, M. & Blackwell, P. Social organization of cooperatively breeding long-tailed tits: kinship and spatial dynamics. *J. Anim. Ecol.*, 820–830 (2001).
19. Okarma, H. *et al*. Home ranges of wolves in Białowieża Primeval Forest, Poland, compared with other Eurasian populations. *J. Mammal.* **79**, 842–852 (1998).
20. Burgman, M. A. & Fox, J. C. in *Animal Conservation Forum*. 19–28 (Cambridge University Press).
21. Office for Coastal Management. Coastal Zone Management Act Boundary. *National Oceanic and Atmospheric Administration* https://www.fisheries.noaa.gov/inport/item/53132 (2024).
22. Brown, E. A., Hellenthal, R. A., Mahon, M. B., Rumschlag, S. L. & Rohr, J. R. Benthic Macroinvertebrate Range Maps. *Figshare* https://doi.org/10.6084/m9.figshare.25517416 (2024).
23. Brown, E. A., Hellenthal, R. A., Mahon, M. B., Rumschlag, S. L. & Rohr, J. R. Benthic Macroinvertebrate Range Map Shapefiles. *Figshare* https://doi.org/10.6084/m9.figshare.25517437 (2024).
24. Brown, E. A., Hellenthal, R. A., Mahon, M. B., Rumschlag, S. L. & Rohr, J. R. Filtered NARS Occurrence Data for Range Maps. *Figshare* https://doi.org/10.6084/m9.figshare.25517455 (2024).
25. Brown, E. A., Hellenthal, R. A., Mahon, M. B., Rumschlag, S. L. & Rohr, J. R. Benthic Macroinvertebrate Waterbody Occupancy Data. *Figshare* https://doi.org/10.6084/m9.figshare.25517470 (2024).
26. Brown, E. A., Hellenthal, R. A., Mahon, M. B., Rumschlag, S. L. & Rohr, J. R. R Code to Generate Benthic Macroinvertebrate Range Maps. *Figshare* https://doi.org/10.6084/m9.figshare.25517494 (2024).
27. Brown, E. A., Hellenthal, R. A., Mahon, M. B., Rumschlag, S. L. & Rohr, J. R. Taxonomic Classification Data for Benthic Macroinvertebrates. *Figshare* https://doi.org/10.6084/m9.figshare.25517482 (2024).
28. Brown, E. A., Hellenthal, R. A., Mahon, M. B., Rumschlag, S. L. & Rohr, J. R. Metadata Document for Range Maps and Other Supplemental Materials. *Figshare* https://doi.org/10.6084/m9.figshare.25517485 (2024).
29. Global Biodiversity Information Facility (GBIF). What is GBIF? https://www.gbif.org/what-is-gbif (2023).
30. Schoch C. L *et al*. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*, https://doi.org/10.1093/database/baaa062 (2020).
31. Shaw, C. A. in *Navigating the Shoals: Evolving User Services in Aquatic and Marine Science Libraries: Proceedings of the 29th Annual Conference of the International Association of Aquatic and Marine Science Libraries and Information Centers (IAMSLIC)*. 17 (IAMSLIC).
32. Chamberlain, S. *et al*. Package 'taxize': Taxonomic Information from Around the Web. *CRAN* https://cran.r-project.org/web/packages/taxize/index.html (2017).
33. Bank, R. A. *et al*. MolluscaBase–announcing a World Register of all Molluscs. (2014).
34. Costello, M. J. *et al*. Global coordination and standardisation in marine biodiversity through the World Register of Marine Species (WoRMS) and related databases. *Plos One* **8**, e51629 (2013).
35. Pebesma, E. *et al*. Package 'sp': Classes and methods for spatial data in R. *CRAN* https://cran.r-project.org/web/packages/sp/index.html (2024).
36. Pebesma, E. *et al*. Package 'sf': Simple Features for R. *CRAN* https://cran.r-project.org/web/packages/sf/index.html (2024).
37. Hijmans, R. J. *et al*. Package 'Raster': Geographic data analysis and modeling. *CRAN* https://cran.r-project.org/web/packages/raster/index.html (2023).
38. R Development Team. R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2022).
39. RStudio Development Team. RStudio: Integrated Development for R v. 2021.9.2.382 (RStudio, PBC, Boston, MA, 2022).

Acknowledgements

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. government. The findings and conclusions here do not necessarily represent the views or policies of the U.S. Environmental Protection Agency. We are indebted to the federal, state, and tribal biologists and contractors who planned the USEPA NARS surveys, collected the samples in the field, and processed and identified samples. This

work was conducted as part of the Analyses of Contaminant Effects in Freshwater Systems: Synthesizing Abiotic and Biotic Stream Datasets for Long-Term Ecological Research Working Group supported by the John Wesley Powell Center for Analysis and Synthesis, funded by the U.S. Geological Survey.

Author contributions

All authors helped to conceptualize and plan the project. E.A.B. cleaned and compiled the NLA and NCCA data, prepared the mapping code, and generated all maps and shapefiles. S.L.R. and M.B.M. cleaned and compiled all NRSA data and provided guidance and insights regarding the use of NARS data. R.A.H. updated the taxonomic classification info based on the most current groupings and provided valuable input on map presentation and organization. E.A.B. wrote the manuscript and created figures. All authors provided feedback and revisions to the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03845-5.

Correspondence and requests for materials should be addressed to E.A.B.