

Bryn Mawr College

## Scholarship, Research, and Creative Work at Bryn Mawr College

---

Biology Faculty Research and Scholarship

Biology

---

2022

### Standardized NEON organismal data for biodiversity research

Daijiang Li

*Louisiana State University at Baton Rouge*

Sydne Record

*Bryn Mawr College, srecord@brynmawr.edu*

Eric R. Sokol

*University of Colorado Boulder*

Matthew E. Bitters

*University of Colorado Boulder*

Melissa Y. Chen

*University of Colorado Boulder*

*See next page for additional authors*

Follow this and additional works at: [https://repository.brynmawr.edu/bio\\_pubs](https://repository.brynmawr.edu/bio_pubs)



Part of the [Biology Commons](#)



This work is licensed under a [Creative Commons Attribution 4.0 International License](#).

[Let us know how access to this document benefits you.](#)

---

#### Citation

Li, D., Record, S., Sokol, E. R., Bitters, M. E., Chen, M. Y., Chung, Y. A., Helmus, M. R., et al. 2022. "Standardized NEON Organismal Data for Biodiversity Research." *Ecosphere* 13.7: e4141.

This paper is posted at Scholarship, Research, and Creative Work at Bryn Mawr College.

[https://repository.brynmawr.edu/bio\\_pubs/54](https://repository.brynmawr.edu/bio_pubs/54)

For more information, please contact [repository@brynmawr.edu](mailto:repository@brynmawr.edu).

---
























## Authors

Daijiang Li, Sydne Record, Eric R. Sokol, Matthew E. Bitters, Melissa Y. Chen, Y. Anny Chung, Matthew R. Helmus, Ruvi Jaimes, Lara Jansen, Marta A. Jarzyna, Michael G. Just, Jalene M. LaMontagne, Brett A. Melbourne, Wynne Moss, Kari E. A. Norman, Stephanie M. Parker, Natalie Robinson, Bijan Seyednasrollah, Colin Smith, Sarah Spaulding, Thilina D. Surasinghe, Sarah K. Thomsen, and Phoebe L. Zarnetske

## ARTICLE

## Special Feature: Harnessing the NEON Data Revolution

# Standardized NEON organismal data for biodiversity research

**Daijiang Li**<sup>1,2</sup>  | **Sydne Record**<sup>3,4</sup>  | **Eric R. Sokol**<sup>5,6</sup>  | **Matthew E. Bitters**<sup>7</sup>  |  
**Melissa Y. Chen**<sup>7</sup>  | **Y. Anny Chung**<sup>8</sup>  | **Matthew R. Helmus**<sup>9</sup>  |  
**Ruvi Jaimes**<sup>10</sup>  | **Lara Jansen**<sup>11</sup>  | **Marta A. Jarzyna**<sup>12,13</sup>  | **Michael G. Just**<sup>14</sup>  |  
**Jalene M. LaMontagne**<sup>15</sup>  | **Brett A. Melbourne**<sup>7</sup>  | **Wynne Moss**<sup>7</sup>  |  
**Kari E. A. Norman**<sup>16</sup>  | **Stephanie M. Parker**<sup>5</sup>  | **Natalie Robinson**<sup>5</sup>  |  
**Bijan Seyednasrollah**<sup>17</sup>  | **Colin Smith**<sup>18</sup>  | **Sarah Spaulding**<sup>6</sup>  |  
**Thilina D. Surasinghe**<sup>19</sup>  | **Sarah K. Thomsen**<sup>20</sup>  | **Phoebe L. Zarnetske**<sup>21,22</sup> 

<sup>1</sup>Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, USA

<sup>2</sup>Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana, USA

<sup>3</sup>Department of Biology, Bryn Mawr College, Bryn Mawr, Pennsylvania, USA

<sup>4</sup>Department of Wildlife, Fisheries, and Conservation Biology, University of Maine, Orono, Maine, USA

<sup>5</sup>National Ecological Observatory Network (NEON), Battelle, Boulder, Colorado, USA

<sup>6</sup>Institute of Arctic and Alpine Research (INSTAAR), University of Colorado Boulder, Boulder, Colorado, USA

<sup>7</sup>Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, Colorado, USA

<sup>8</sup>Departments of Plant Biology and Plant Pathology, University of Georgia, Athens, Georgia, USA

<sup>9</sup>Integrative Ecology Lab, Center for Biodiversity, Department of Biology, Temple University, Philadelphia, Pennsylvania, USA

<sup>10</sup>St. Edward's University, Austin, Texas, USA

<sup>11</sup>Department of Environmental Science and Management, Portland State University, Portland, Oregon, USA

<sup>12</sup>Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, Ohio, USA

<sup>13</sup>Translational Data Analytics Institute, The Ohio State University, Columbus, Ohio, USA

<sup>14</sup>Ecological Processes Branch, U.S. Army ERDC CERL, Champaign, Illinois, USA

<sup>15</sup>Department of Biological Sciences, DePaul University, Chicago, Illinois, USA

<sup>16</sup>Department of Environmental Science, Policy, and Management, University of California Berkeley, Berkeley, California, USA

<sup>17</sup>School of Informatics, Computing and Cyber Systems, Northern Arizona University, Flagstaff, Arizona, USA

<sup>18</sup>Environmental Data Initiative, University of Wisconsin-Madison, Madison, Wisconsin, USA

<sup>19</sup>Department of Biological Sciences, Bridgewater State University, Bridgewater, Massachusetts, USA

<sup>20</sup>Department of Integrative Biology, Oregon State University, Corvallis, Oregon, USA

<sup>21</sup>Department of Integrative Biology, Michigan State University, East Lansing, Michigan, USA

<sup>22</sup>Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, Michigan, USA

Daijiang Li, Sydne Record, and Eric Sokol contributed equally to the work reported here.

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.

© 2022 The Authors. *Ecosphere* published by Wiley Periodicals LLC on behalf of The Ecological Society of America.

**Correspondence**

Daijiang Li

Email: [dli30@lsu.edu](mailto:dli30@lsu.edu)**Funding information**

National Science Foundation, Grant/  
Award Numbers: 1832282, 1906144,  
1926341, 1926567, 1926568, 1926598

**Handling Editor:** Jennifer K. Balch**Abstract**

Understanding patterns and drivers of species distribution and abundance, and thus biodiversity, is a core goal of ecology. Despite advances in recent decades, research into these patterns and processes is currently limited by a lack of standardized, high-quality, empirical data that span large spatial scales and long time periods. The NEON fills this gap by providing freely available observational data that are generated during robust and consistent organismal sampling of several sentinel taxonomic groups within 81 sites distributed across the United States and will be collected for at least 30 years. The breadth and scope of these data provide a unique resource for advancing biodiversity research. To maximize the potential of this opportunity, however, it is critical that NEON data be maximally accessible and easily integrated into investigators' workflows and analyses. To facilitate its use for biodiversity research and synthesis, we created a workflow to process and format NEON organismal data into the ecomDP (ecological community data design pattern) format that were available through the ecomDP R package; we then provided the standardized data as an R data package (neonDivData). We briefly summarize sampling designs and data wrangling decisions for the major taxonomic groups included in this effort. Our workflows are open-source so the biodiversity community may: add additional taxonomic groups; modify the workflow to produce datasets appropriate for their own analytical needs; and regularly update the data packages as more observations become available. Finally, we provide two simple examples of how the standardized data may be used for biodiversity research. By providing a standardized data package, we hope to enhance the utility of NEON organismal data in advancing biodiversity research and encourage the use of the harmonized ecomDP data design pattern for community ecology data from other ecological observatory networks.

**KEYWORDS**

biodiversity, data package, data product, EDI, organismal data, R, Special Feature: Harnessing the NEON Data Revolution

## INTRODUCTION (OR WHY STANDARDIZED NEON ORGANISMAL DATA)

A central goal of ecology is to understand the patterns and processes of biodiversity, and this is particularly important in an era of rapid global environmental change (Blowes et al., 2019; Midgley & Thuiller, 2005). Such understanding is only possible through studies that address questions such as: How is biodiversity distributed across large spatial scales, ranging from ecoregions to continents? What mechanisms drive spatial patterns of biodiversity? Are spatial patterns of biodiversity similar among different taxonomic groups, and if not, why do we

see variation? How does community composition vary across spatial and environmental gradients? What are the local- and landscape-scale drivers of community structure? How and why do biodiversity patterns change over time? Answers to such questions will enable better management and conservation of biodiversity and ecosystem services.

Biodiversity research has a long history (Worm & Tittensor, 2018), beginning with major scientific expeditions (e.g., Alexander von Humboldt, Charles Darwin) aiming to document global species lists after the establishment of Linnaeus's *Systema Naturae* (Linnaeus, 1758). Beginning in the 1950s (Curtis, 1959; Hutchinson, 1959), researchers moved beyond documentation to focus on

quantifying patterns of species diversity and describing mechanisms underlying their heterogeneity. Since the beginning of this line of research, major theoretical breakthroughs (Brown et al., 2004; Harte, 2011; Hubbell, 2001; MacArthur & Wilson, 1967) have advanced our understanding of potential mechanisms causing and maintaining biodiversity. Modern empirical studies, however, have been largely constrained to local or regional scales and focused on one or a few taxonomic groups, because of the considerable effort required to collect observational data. There are now unprecedented numbers of observations from independent small and short-term ecological studies. These data support research into generalities through syntheses and meta-analyses (Blowes et al., 2019; Li et al., 2020; Vellend et al., 2013), but this work is challenged by the difficulty of integrating data from different studies and with varying limitations. Such limitations include the following: differing collection methods (methodological uncertainties); varying levels of statistical robustness; inconsistent handling of missing data; spatial bias; publication bias; and design flaws (Koricheva & Gurevitch, 2014; Martin et al., 2012; Nakagawa & Santos, 2012; Welte et al., 2021). Additionally, it has historically been challenging for researchers to obtain and collate data from a diversity of sources for use in syntheses and/or meta-analyses (Gurevitch & Hedges, 1999).

Barriers to meta-analyses have been reduced in recent years to bring biodiversity research into the big data era (Farley et al., 2018; Hampton et al., 2013) by large efforts to digitize museum and herbarium specimens (e.g., iDigBio), successful community science programs (e.g., iNaturalist, eBird), technological advances (e.g., remote sensing, automated acoustic recorders), and long-running coordinated research networks. Yet, each of these remedies comes with its own limitations. For instance, museum/herbarium specimens and community science records are increasingly available, but are still incidental and unstructured in terms of the sampling design, and exhibit marked geographic and taxonomic biases (Beck et al., 2014; Geldmann et al., 2016; Martin et al., 2012). Remote sensing approaches may cover large spatial scales, but may also be of low spatial resolution and unable to reliably penetrate vegetation canopy (Palumbo et al., 2017; Pricope et al., 2019). The standardized observational sampling of woody trees by the US Forest Service's Forest Inventory and Analysis and of birds by the US Geological Survey's Breeding Bird Survey has been ongoing across the United States since 2001 and 1966, respectively (Bechtold & Patterson, 2005; Sauer et al., 2017), but covers few taxonomic groups. The Long Term Ecological Research Network (LTER) and Critical Zone Observatory (CZO) both are hypothesis-driven research efforts built on decades of previous work (Jones

et al., 2021). While both provide considerable observational and experimental datasets for diverse ecosystems and taxa, their sampling and dataset design are tailored to their specific research questions and a priori standardization is not possible. Thus, despite recent advances, biodiversity research is still impeded by a lack of standardized, high-quality, and open-access data spanning large spatial scales and long time periods.

The recently established NEON provides continental-scale observational and instrumentation data for a wide variety of taxonomic groups and measurement streams. Data are collected using standardized methods, across 81 field sites in both terrestrial and freshwater ecosystems, and will be freely available for at least 30 years. These consistently collected, long-term, and spatially robust measurements are directly comparable throughout the observatory, and provide a unique opportunity for enabling a better understanding of ecosystem change and biodiversity patterns and processes across space and through time (Keller et al., 2008).

NEON data are designed to be maximally useful to ecologists by aligning with FAIR principles (findable, accessible, interoperable, and reusable; Wilkinson et al., 2016). Despite meeting these requirements, however, there are still challenges to integrating NEON organismal data (e.g., occurrence and abundance of species) for reproducible biodiversity research. For example, field names may vary across NEON data products, even for similar measurements; some measurements include sampling unit information, whereas units must be decided for others. These issues and inconsistencies may be overcome through data cleaning and formatting, but understanding how best to perform this task requires a significant investment in the comprehensive NEON documentation for each data product involved in an analysis. Thoroughly reading large amounts of NEON documentation is time-consuming, and the path to a standard data format, as is critical for reproducibility, may vary greatly between NEON organismal data products and users—even for similar analyses. Ultimately, this may result in subtle differences from study to study that hinder meta-analyses using NEON data. A simplified and standardized format for NEON organismal data would facilitate wider usage of these datasets for biodiversity research. Furthermore, if these data were formatted to interface well with datasets from other coordinated research networks, more comprehensive syntheses could be accomplished to advance macrosystems biology (Record et al., 2020).

One attractive standardized formatting style for NEON organismal data is that of *ecocomDP* (ecological community data design pattern; O'Brien et al., 2021). *EcocomDP* is the brainchild of members of the LTER network, the Environmental Data Initiative (EDI), and

NEON staff, and provides a model by which data from a variety of sources may be easily transformed into consistently formatted, analysis-ready community-level organismal data packages. This is done using reproducible code that maintains dataset “levels”: L0 is incoming data, L1 represents an ecomDP data format and includes tables representing observations, sampling locations, and taxonomic information (at a minimum), and L2 is an output format. Thus far, >70 LTER organismal datasets have been harmonized to the L1 ecomDP format through the R package ecomDP and more datasets are in the queue for processing into the ecomDP format by EDI (O’Brien et al., 2021).

We standardized NEON organismal data into the ecomDP format, and all R codes to process NEON data products can be obtained through the R package ecomDP. For the major taxonomic groups included in this initial effort, NEON sampling designs and major data wrangling decisions are summarized in the *Materials and Methods* section. We archived the standardized data in the EDI Data Repository (<https://doi.org/10.6073/pasta/c28dd4f6e7989003505ea02e9a92afbf>). To facilitate the usage of the standardized datasets, we also developed an R data package, *neonDivData* (<https://github.com/daijiang/neonDivData>). We refer to the input data streams provided by NEON as data products, and the cleaned and standardized collection of data files provided here as objects within the R data package, *neonDivData*, across this paper. Standardized datasets will be maintained and updated as new data become available from the NEON portal. We hope this effort will substantially reduce data processing times for NEON data users and greatly facilitate the use of NEON organismal data to advance our understanding of Earth’s biodiversity.

## MATERIALS AND METHODS (OR HOW TO STANDARDIZE NEON ORGANISMAL DATA)

There are many details to consider when starting to use NEON organismal data products. Below, we outline key points relevant to community-level biodiversity analyses with regard to the NEON sampling design and decisions that were made as the data products presented in this paper were converted into the ecomDP data model. While the methodological sections below are specific to particular taxonomic groups, there are some general points that apply to all NEON organismal data products. First, species occurrence and abundance measures as reported in NEON biodiversity data products are not standardized to sampling effort. Because there are often multiple approaches to cleaning (e.g., dealing with multiple levels

of taxonomic resolution, interpretations of absences) and standardizing biodiversity survey data, NEON publishes raw observations along with sampling effort data to preserve as much information as possible so that data users can clean and standardize data as they see fit. The workflows described here for 12 taxonomic groups represented in 11 NEON data products produce standardized counts based on sampling effort, such as the count of individuals per area sampled or count standardized to the duration of trap deployment, as described in Table 1. The data wrangling workflows described below can be used to access, download, and clean data from the NEON Data Portal using the R ecomDP package. To view a catalog of available NEON data products in the ecomDP format, use `ecomDP::search_data` (“NEON”). To import data from a given NEON data product into your R environment, use `ecomDP::read_data()`, and set the `id` argument to the selected NEON to ecomDP mapping workflow (the “L0 to L1 ecomDP workflow ID” in Table 1). This will return a list of ecomDP formatted tables and accompanying metadata. To create a flat data table (similar to the R objects in the data package *neonDivData* described in Table 2), use the `ecomDP::flatten_data()` function.

Second, because different taxonomic groups have different sampling designs (see below for details), there are no general data processing protocol that can be applied to all taxonomic groups. Nevertheless, we tried to be as consistent as possible during the data cleaning and standardization processes. All final data products have the minimal information of locations (e.g., `location_id`, `site_id`, `plot_id`), species names (e.g., `taxon_id`, `taxon_name`, `taxon_rank`), and presence/absence or abundance information (e.g., `variable_name`, `value`, `unit`).

Third, our processes assume that NEON ensured correct identifications of species. However, since records may be identified to any level of taxonomic resolution, and IDs above the genus level may not be useful for most biodiversity projects, we removed records with such IDs for groups that are relatively easy to identify (i.e., fish, plant, small mammals) or have very few taxon IDs that are above genus level (i.e., mosquito). However, for groups that are hard to identify (i.e., algae, beetle, bird, macroinvertebrate, tick, and tick pathogen), we decided to keep all records regardless of their taxon ID level. Users thus need to carefully consider which level of taxon IDs they need to address their research questions. Another note regarding species names is the term “sp.” versus “spp.” across NEON organismal data collections; the term “sp.” refers to a single morphospecies, whereas the term “spp.” refers to more than one morphospecies. This is an important point to consider for community ecology or biodiversity analyses because it may add uncertainty to estimates of biodiversity



**TABLE 1** Mapping NEON data products to ecomDP formatted data packages with abundance standardized to observation effort.

| Taxon group                 | L0 dataset (NEON data product ID) | Version of NEON data used in this study   | L0 to L1 ecomDP workflow ID | Primary variable reported in ecomDP observation table | Units   |
|-----------------------------|-----------------------------------|---|-----------------------------|---|---|
| Algae                       | DP1.20166.001                     | <a href="https://doi.org/10.48443/g2k4-d258">https://doi.org/10.48443/g2k4-d258</a> ; and provisional data  | neon.ecocomdp.20166.001.001 | Cell density  | Cells per square centimeter or cells per milliliter       |
| Beetles                     | DP1.10022.001                     | <a href="https://doi.org/10.48443/xgea-hw23">https://doi.org/10.48443/xgea-hw23</a> ; and provisional data  | neon.ecocomdp.10022.001.001 | Abundance   | Count per trap day  |
| Birds <sup>a</sup>          | DP1.10003.001                     | <a href="https://doi.org/10.48443/88sy-ah40">https://doi.org/10.48443/88sy-ah40</a> ; and provisional data  | neon.ecocomdp.10003.001.001 | Cluster size  | Count of individuals                                      |
| Fish                        | DP1.20107.001                     | <a href="https://doi.org/10.48443/7p84-6j62">https://doi.org/10.48443/7p84-6j62</a> ; and provisional data  | neon.ecocomdp.20107.001.001 | Abundance   | Catch per unit effort                                     |
| Herptiles                   | DP1.10022.001                     | <a href="https://doi.org/10.48443/xgea-hw23">https://doi.org/10.48443/xgea-hw23</a> ; and provisional data  | neon.ecocomdp.10022.001.002 | Abundance   | Count per trap day  |
| Macroinvertebrates          | DP1.20120.001                     | <a href="https://doi.org/10.48443/gn8x-k322">https://doi.org/10.48443/gn8x-k322</a> ; and provisional data  | neon.ecocomdp.20120.001.001 | Density   | Count per square meter                                    |
| Mosquitoes                  | DP1.10043.001                     | <a href="https://doi.org/10.48443/c7h7-q918">https://doi.org/10.48443/c7h7-q918</a> ; and provisional data  | neon.ecocomdp.10043.001.001 | Abundance   | Count per trap hour                                       |
| Plants <sup>b</sup>         | DP1.10058.001                     | <a href="https://doi.org/10.48443/abge-r811&amp;#12">https://doi.org/10.48443/abge-r811&amp;#12</a><br><a href="https://doi.org/10.48443/pr5e-1q60">https://doi.org/10.48443/pr5e-1q60</a> ; and provisional data | neon.ecocomdp.10058.001.001 | Percent cover   | Percent of plot area covered by taxon                     |
| Small mammals               | DP1.10072.001                     | <a href="https://doi.org/10.48443/h3dk-3a71">https://doi.org/10.48443/h3dk-3a71</a> ; and provisional data  | neon.ecocomdp.10072.001.001 | Count   | Unique individuals per 100 trap nights per plot per month |
| Tick pathogens <sup>c</sup> | DP1.10092.001                     | <a href="https://doi.org/10.48443/nygx-dm71">https://doi.org/10.48443/nygx-dm71</a> ; and provisional data  | neon.ecocomdp.10092.001.001 | Positivity rate                                       | Positive tests per pathogen per sampling event            |
| Ticks                       | DP1.10093.001                     | <a href="https://doi.org/10.48443/7jh5-8s51">https://doi.org/10.48443/7jh5-8s51</a> ; and provisional data  | neon.ecocomdp.10093.001.001 | Abundance   | Count per square meter                                    |
| Zooplankton                 | DP1.20219.001                     | <a href="https://doi.org/10.48443/150d-yf27">https://doi.org/10.48443/150d-yf27</a> ; and provisional data  | neon.ecocomdp.20219.001.001 | Density   | Count per liter   |

Note: IDs in the L0 to L1 ecomDP workflow ID columns were used in the R package ecomDP to standardize organismal data.

<sup>a</sup>Bird counts are reported per taxon per “cluster” observed in each point count in the NEON data product and have not been further standardized to sampling effort because standard methods for modeling bird abundance are beyond the scope of this paper.

<sup>b</sup>Plants’ percent cover value NA represents presence/absence data only.

<sup>c</sup>Incidence rate per number of tests conducted is reported for tick pathogens.

metrics such as species richness. It is also important to point out that NEON fuzzed taxonomic IDs to one higher taxonomic level to protect species of concern. For example, if a threatened Black-capped vireo (*Vireo*

*atricapilla*) is recorded by a NEON technician, the taxonomic identification is fuzzed to *Vireo* in the data. Rare, threatened, and endangered species are those listed as such by federal and/or state agencies.

**TABLE 2** Summary of data products included in this study (as of 13 April 2022). Users can call the R objects in the R object column from the R data package `neonDivData` to get the standardized data for specific taxonomic groups.

| Taxon group        | R object                            | No. species | No. sites | Start date    | End date         |
|--------------------|-------------------------------------|-------------|-----------|---------------|------------------|
| Algae              | <code>data_algae</code>             | 2279        | 34        | 2 July 2014   | 28 July 2020     |
| Beetles            | <code>data_beetle</code>            | 768         | 47        | 3 July 2013   | 5 October 2021   |
| Birds              | <code>data_bird</code>              | 577         | 47        | 5 June 2013   | 16 July 2021     |
| Fish               | <code>data_fish</code>              | 158         | 28        | 29 March 2016 | 14 December 2021 |
| Herptiles          | <code>data_herp_bycatch</code>      | 136         | 41        | 2 April 2014  | 6 October 2021   |
| Macroinvertebrates | <code>data_macroinvertebrate</code> | 1373        | 34        | 1 July 2014   | 28 July 2021     |
| Mosquitoes         | <code>data_mosquito</code>          | 131         | 47        | 9 April 2014  | 11 June 2021     |
| Plants             | <code>data_plant</code>             | 6544        | 47        | 24 June 2013  | 13 October 2021  |
| Small mammals      | <code>data_small_mammal</code>      | 149         | 46        | 19 June 2013  | 18 November 2021 |
| Tick pathogens     | <code>data_tick_pathogen</code>     | 12          | 16        | 17 April 2014 | 1 October 2020   |
| Ticks              | <code>data_tick</code>              | 19          | 46        | 2 April 2014  | 18 November 2021 |
| Zooplankton        | <code>data_zooplankton</code>       | 166         | 7         | 2 July 2014   | 21 July 2021     |

Fourth, NEON publishes data for additional organismal groups, which were not included in this study given the complexity of the data. For example, aquatic plants (DP1.20066.001 and DP1.20072.001); benthic microbe abundance (DP1.20277.001), metagenome sequences (DP1.20279.001), marker gene sequences (DP1.20280.001), and community composition (DP1.20086.001); surface water microbe abundance (DP1.20278.001), metagenome sequences (DP1.20281.001), marker gene sequences (DP1.20282.001), and community composition (DP1.20141.001); and soil microbe biomass (DP1.10104.001), metagenome sequences (DP1.10107.001), marker gene sequences (DP1.10108.001), and community composition (DP1.10081.001) were not considered here, though future work may utilize `neonDivData` to align these datasets. Users interested in further explorations of these data products may find more information on the NEON data portal (<https://data.neonscience.org/>). Additionally, concurrent work on a suggested bioinformatics pipeline and how to run sensitivity analyses on user-defined parameters for NEON soil microbial data, including code and vignettes, is described in Qin et al. (2021).

Finally, it should be noted that NEON data collection efforts will continue well after this paper is published and new changes to data collection methods and/or processing may vary over time. Such changes (e.g., change in the number of traps used for ground beetle collection) or interruptions (e.g., due to COVID-19) to data collection are documented in the issue log for each data product on the NEON Data Portal and the Readme text file that is included with NEON data downloads. We will try our best to maintain and update our standardized data products as long as possible.

## Terrestrial organisms

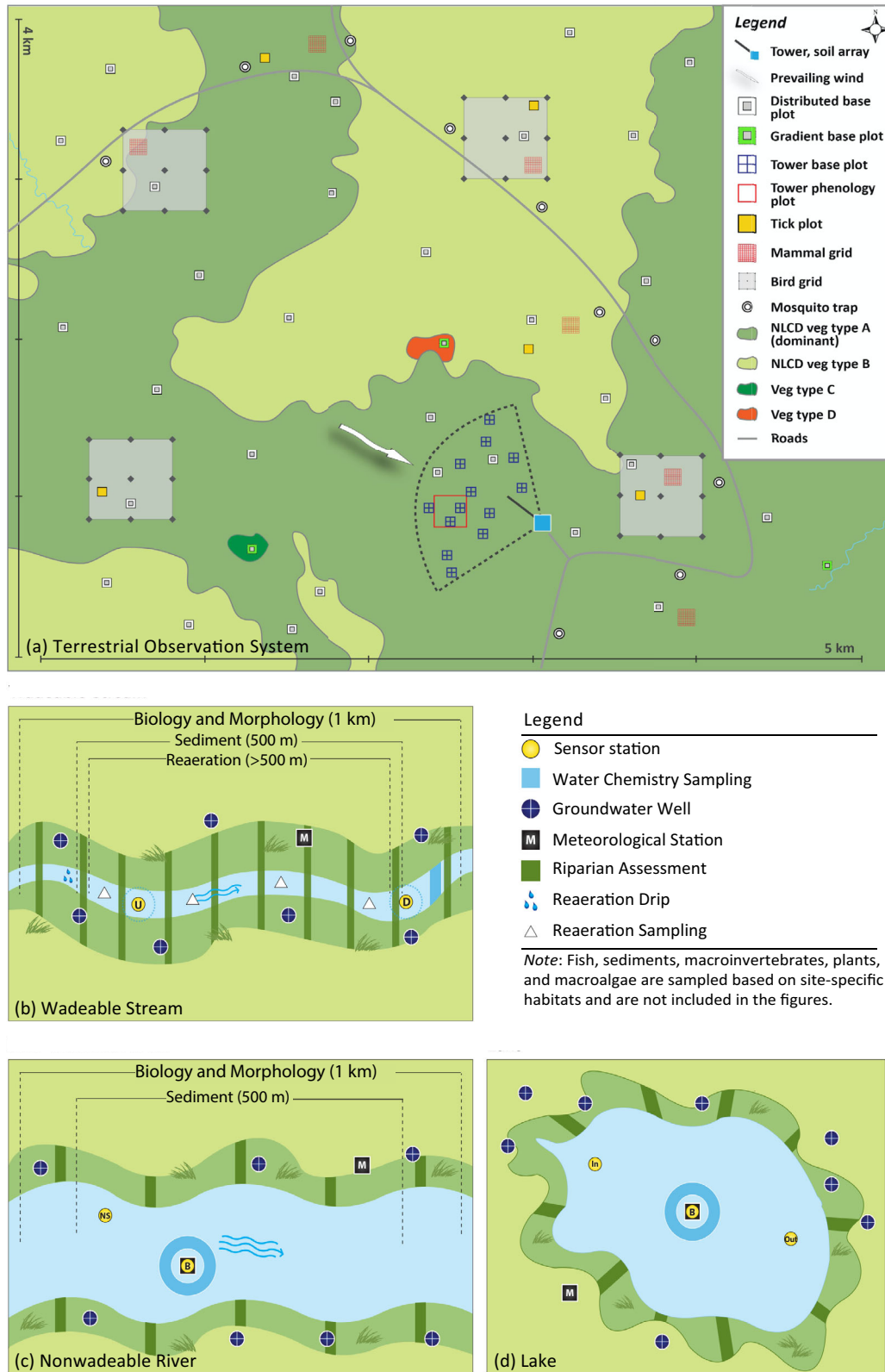
### Breeding landbirds

#### *NEON sampling design*

NEON designates breeding landbirds as “smaller birds (usually exclusive of raptors and upland game birds) not usually associated with aquatic habitats” (Ralph, 1993; Thibault, 2018). Most species observed are diurnal and include both resident and migrant species. Landbirds are surveyed via point counts in each of the 47 terrestrial sites (Thibault, 2018). At most NEON sites, breeding landbird points are located in 5–10, 3 × 3 grids (Figure 1), which are themselves located in representative (dominant) vegetation. Whenever possible, grid centers are co-located with distributed base plot centers. When sites are too small to support a minimum of five grids, separated by at least 250 m from edge to edge, point counts are completed at single points instead of grids. In these cases, points are located at the southwest corners of distributed base plots within the site. Five to 25 points may be surveyed depending on the size and spatial layout of the site, with exact point locations dictated by a stratified-random spatial design that maintains a 250-m minimum separation between points.

Surveys occur during one or two sampling bouts per season, at large and small sites, respectively. Observers go to the specified points early in the morning and track birds observed during each minute of a 6-min period, following a 2-min acclimation period, at each point (Thibault, 2018). Each point count contains species, sex, and distance to each bird (measured with a laser rangefinder except in the case of flyovers) seen or heard.





**FIGURE 1** Generalized sampling schematics for Terrestrial Observation System (TOS) (a) and Aquatic Observation System (B–D) plots. For TOS plots, distributed, tower, and gradient plots, and locations of various sampling regimes are presented via symbols. For Aquatic Observation System plots, wadeable streams, nonwadeable streams, and lake plots are shown in detail, with locations of sensors and different sampling regimes presented using symbols. Panel (a) was originally published in Thorpe et al. (2016).

Information relevant for subsequent modeling of detectability is also collected during the point counts (e.g., weather, detection method). The point count surveys for NEON were modified from the Integrated Monitoring in Bird Conservation Regions field protocol for spatially balanced sampling of landbird populations (Pavlacky Jr et al., 2017).

#### *Data wrangling decisions*

The bird point count NEON data product (“DP1.10003.001”) consists of a list of two associated data frames: `brd_countdata` and `brd_perpoint`. The former data frame contains information such as locations, species identities, and their counts. The latter data frame contains additional location information such as latitude and longitude coordinates and environmental conditions during the time of the observations. The separate data frames are linked by “eventID,” which refers to the location, date, and time of the observation. To prepare the bird point count data for the L1 ecomDP model, we first merged both data frames into one and then removed columns that are likely not needed for most community-level biodiversity analyses (e.g., observer names). The field `taxon_id` in the R object `data_bird` with the `neonDivData` data package consists of the standard AOU four-letter species code, although `taxon_rank` refers to seven potential levels of identification (class, family, genus, species, speciesGroup, subfamily, and subspecies). Users can decide which level is appropriate; for example, one might choose to exclude all unidentified birds (`taxon_id = UNBI`), where no further details are available below the class level (*Aves* sp.). The NEON sampling protocol has evolved over time, so users are advised to check whether the “samplingProtocolVersion” associated with bird point count data (“DP1.10003.001”) fits their data requirements and subset as necessary. Older versions of protocols can be found at the NEON document library.

## Ground beetle and herp bycatch

#### *NEON sampling design*

Ground beetle sampling is conducted via pitfall trapping, across 10 distributed plots at each NEON site. The original sampling design included the placement of a pitfall trap at each of the cardinal directions along the distributed plot boundary, for a total of four traps per plot and 40 traps per site. In 2018, sampling was reduced via the elimination of the North pitfall trap in each plot, resulting in 30 traps per site (LeVan, Robinson, et al., 2019).

Beetle pitfall trapping begins when the temperature has been  $>4^{\circ}\text{C}$  for 10 days in the spring and ends when temperatures dip below this threshold in the fall.

Sampling occurs biweekly throughout the sampling season with no single trap being sampled more frequently than every 12 days (LeVan, 2020a). After collection, the samples are separated into carabid species and bycatch.

Invertebrate bycatch is pooled to the plot level and archived. Vertebrate bycatch is sorted and identified by NEON technicians, then archived at the trap level. Carabid samples are sorted and identified by NEON technicians, after which a subset of carabid individuals are sent to be pinned and reidentified by an expert taxonomist. More details can be found in Hoekman et al. (2017) and LeVan, Robinson, et al. (2019).

Pitfall traps and sampling methods are designed by NEON to reduce vertebrate bycatch (LeVan, Robinson, et al., 2019). The pitfall cup is medium in size with a low clearance cover installed over the trap entrance to minimize large vertebrate bycatch. When a live vertebrate with the ability to move on its own volition is found in a trap, the animal is released. Live but moribund vertebrates are euthanized and collected along with deceased vertebrates. When  $\geq 15$  individuals of a vertebrate species are collected, cumulatively, within a single plot, NEON may initiate localized mitigation measures such as temporarily deactivating traps and removing all traps from the site for the remainder of the season. Thus, while herpetofaunal (herp) bycatch is present in many pitfall samples it is unclear how well these pitfall traps capture herp community structure and diversity—due to these active efforts to reduce vertebrate bycatch. Users of NEON herp bycatch data should be aware of these limitations.

#### *Data wrangling decisions*

The beetle and herp bycatch data product identifier is “DDP1.10022.001.” Carabid samples are recorded and identified in a multistep workflow wherein a subset of samples are passed on in each successive step. Individuals are first identified by the sorting technician after which a subset is sent on to be pinned. Some especially difficult individuals are not identified by technicians during sorting, instead of being labeled “other carabid.” The identifications for those individuals are recorded with the pinning data. Any individuals for which identification is still uncertain are then verified by an expert taxonomist. There are a few cases where an especially difficult identification was sent to multiple expert taxonomists, and they did not agree on a final taxon; these individuals were excluded from the dataset at the recommendation of NEON staff.

Preference is given to expert identification whenever available. However, these differences in taxonomic expertise do not seem to cause systematic biases in estimating species richness across sites, but nonexpert taxonomists are more likely to misidentify non-native carabid species (Egli

et al., 2020). Beetle abundance is recorded for the sorted samples by NEON technicians. To account for individual samples that were later reidentified, the final abundance for a species is the original sorting sample abundance minus the number of individuals that were given a new ID.

Prior to 2018, trappingDays values were not included for many sites. Missing entries were calculated as the range from setDate through collectDate for each trap. We also accounted for a few plots for which setDate was not updated based on a previous collection event in the trappingDays calculations. To facilitate easy manipulation of data within and across bouts, a new boutID field was created to identify all trap collection events at a site in a bout. The original EventID field is intended to identify a bout, but has a number of issues that necessitates the creation of a new ID. First, EventID does not correspond to a single collection date but rather all collections in a week. This is appropriate for the small number of instances when collections for a bout happen over multiple consecutive days (~5% of bouts), but prevents analysis of bout patterns at the temporal scale of a weekday. The data here were updated so all entries for a bout correspond to the date (i.e., collectDate) on which the majority of traps are collected to maintain the weekday-level resolution with as high of fidelity as possible, while allowing for easy aggregation within bouts and collectDates. Second, there were a few instances in which plots within a site were set and collected on the same day, but have different EventIDs. These instances were all considered a single bout by our new boutID, which is a unique combination of setDate, collectDate, and siteID.

Herpetofaunal bycatch (amphibian and reptile) in pitfall traps were identified to species or the lowest taxonomic level possible within 24 h of recovery from the field. To process the herp bycatch NEON data, we cleaned trappingDays and the other variables and added boutID as described above for beetles. The variable sampleType in the bet\_sorting table provides the type of animal caught in a pitfall trap as one of five types: “carabid,” “vert bycatch herp,” “other carabid,” “invert bycatch,” and “vert bycatch mam.” We filtered the beetle data described above to only include the “carabid” and “other carabid” types. For herps, we only kept the sampleType of “vert bycatch herp.” Abundance data of beetle and herp bycatch were standardized to be the number of individuals captured per trap day.

## Mosquitoes

### *NEON sampling design*

Mosquito specimens are collected at 47 terrestrial sites across all NEON domains, and the data are reported in

NEON data product DP1.10043.001. Traps are distributed throughout each site according to a stratified-random spatial design used for all Terrestrial Observation System sampling, maintaining stratification across dominant (>5% of total cover) vegetation types (LeVan, 2020b). The number of mosquito traps placed in each vegetation type is proportional to its percent cover, until 10 total mosquito traps have been placed in the site. Mosquito traps are typically located within 30 m of a road to facilitate expedient sampling and are placed at least 300 m apart to maintain independence.

Mosquito monitoring is divided into off-season and field season sampling (LeVan, Paull, et al., 2019). Off-season sampling begins after three consecutive zero-catch field sampling bouts have occurred, and represents a reduced sampling regime that is designed for the rapid detection of when the next field season should begin and to provide mosquito phenology data. Off-season sampling is conducted at three dedicated mosquito traps spread throughout each core site, while temperatures are >10°C. Once per week, technicians deploy traps at dusk and then collect them at dawn the following day.

Field season sampling begins when the first mosquito is detected during off-season sampling (LeVan, Paull, et al., 2019). Technicians deploy traps at all 10 dedicated mosquito trap locations per site. Traps remain out for a 24-h period or sampling bout, and bouts occur every 2–4 weeks at core and relocatable terrestrial sites, respectively. During the sampling bout, traps are serviced twice and yield one night-active sample, collected at dawn or about 8 h after the trap was set, and 1 day-active sample, collected at dusk or ~16 h after the trap was set. Thus, a 24-h sampling bout yields 20 samples from 10 traps.

NEON collects mosquito specimens using Center for Disease Control (CDC) CO<sub>2</sub> light traps (LeVan, Paull, et al., 2019). These traps have been used by other public health and mosquito-control agencies for a half-century so that NEON mosquito data align across NEON field sites and with existing long-term datasets. A CDC CO<sub>2</sub> light trap consists of a cylindrical insulated cooler that contains dry ice, a plastic rain cover attached to a battery-powered light/fan assembly, and a mesh collection cup. During deployment, the dry ice sublimates and releases CO<sub>2</sub>. Mosquitoes attracted to the CO<sub>2</sub> bait are sucked into the mesh collection cup by the battery-powered fan, where they remain alive until trap collection.

Following field collection, NEON’s field ecologists process, package, and ship the samples to an external laboratory where mosquitoes are identified to species and sex (when possible). A subset of identified mosquitoes are tested for infection by pathogens to quantify the presence/absence and prevalence of various arboviruses. Some mosquitoes are set aside for DNA barcode analysis and long-

term archiving. Particularly rare or difficult-to-identify mosquito specimens are prioritized for DNA barcoding. More details can be found in LeVan, Paull, et al. (2019).

#### *Data wrangling decisions*

The mosquito data product (DP1.10043.001) consists of four data frames: trapping data (`mos_trapping`), sorting data (`mos_sorting`), archiving data (`mos_archivepooling`), and expert taxonomist processed data (`mos_expert-TaxonomistIDProcessed`). We first removed rows (records) with missing information about location, collection date, and sample or subsample ID for all data frames. We then merged all four data frames into one, wherein we only kept records for target taxa (i.e., `targetTaxaPresent = "Y"`) with no known compromised sampling condition (i.e., `sampleCondition = "No known compromise"`). We further removed a small number of records with species identified only to the family level; all remaining records were identified at least to the genus level. We estimated the total individual count per trap hour for each species within a trap as  $(\text{individualCount}/\text{subsampleWeight}) \times \text{totalWeight}/\text{trapHours}$ . We then removed columns that were not likely to be used for calculating biodiversity values.

## Small mammals

#### *NEON sampling design*

NEON defines small mammals based on taxonomic, behavioral, dietary, and size constraints, and includes any rodent that (1) is nonvolant; (2) is nocturnally active; (3) forages predominantly aboveground; and (4) has a mass  $>5$  g, but  $<500$ – $600$  g (Thibault et al., 2019). In North America, this includes cricetids, heteromyids, small sciurids, and introduced murids, but excludes shrews, large squirrels, rabbits, or weasels, although individuals of these species may be incidentally captured.

Small mammals are collected at NEON sites using Sherman traps, identified to species in the field, marked with a unique tag, and released (Thibault et al., 2019). Multiple  $90 \times 90$  m trapping grids are set up in each terrestrial field site within the dominant vegetation type. Each  $90 \times 90$  m trapping grid contains 100 traps placed in a pattern with 10 rows and 10 columns set 10 m apart. Three of these  $90 \times 90$  m grids per site are designated pathogen (as opposed to diversity) grids, and additional blood sampling is conducted here.

Small mammal sampling occurs in bouts, with a bout comprised of three consecutive (or nearly consecutive) nights of trapping at each pathogen grid and one night of trapping at each diversity grid. The timing of sampling occurs within 10 days before or after the new moon. The

number of bouts per year is determined by site type: Core sites are typically trapped for six bouts per year (except for areas with shorter seasons due to cold weather), while relocatable sites are trapped for four bouts per year. More information can be found in Thibault et al. (2019).

#### *Data wrangling decisions*

In the small mammal NEON data product (DP1.10072.001), records are stratified by NEON site, year, month, and day and represent data from both the diversity and pathogen sampling grids. Capture records were removed if they were not identified to genus or species (e.g., if the species name was denoted as “either/or” or as family name), or if their trap status is not “5—capture” or “4—more than 1 capture in one trap.” Abundance data for each plot and month combination were standardized to be the number of individuals captured per 100 trap nights.

## Terrestrial plants

#### *NEON sampling design*

NEON plant diversity sampling is completed once or twice per year (one or two “bouts”) in multiscale,  $400\text{-m}^2$  ( $20 \times 20$  m) plots (Barnett, 2019). Each multiscale plot is subdivided into four  $100\text{-m}^2$  ( $10 \times 10$  m) subplots that each encompasses one or two sets of  $10\text{-m}^2$  ( $3.16 \times 3.16$  m) subplots within which a  $1\text{-m}^2$  ( $1 \times 1$  m) subplot is nested. The percent cover of each plant species is estimated visually in the  $1\text{-m}^2$  subplots, while only species presences are documented in the  $10$ - and  $100\text{-m}^2$  subplots.

To estimate plant percent cover by species, technicians record this value for all species in a  $1\text{-m}^2$  subplot (Barnett, 2019). Next, the remaining  $9\text{-m}^2$  area of the associated  $10\text{-m}^2$  subplot is searched for the presence of species. The process is repeated if there is a second  $1$ - and  $10\text{-m}^2$  nested pair in the specific  $100\text{-m}^2$  subplot. Next, the remaining  $80\text{-m}^2$  area is searched for the presence of species; data can be aggregated for a complete list of species present at the  $100\text{-m}^2$  subplot scale. Data for all four  $100\text{-m}^2$  subplots represent indices of species at the  $400\text{-m}^2$  plot scale. In most cases, species encountered in a nested, finer scale, subplot are not rerecorded in any corresponding larger subplot—in order to avoid duplication. Plant species are occasionally recorded more than once, however, when data are aggregated across all nested subplots within each  $400\text{-m}^2$  plot, and these require removal from the dataset. More details about the sampling design can be found in Barnett et al. (2019).

NEON manages plant taxonomic entries with a master taxonomy list that is based on the community standard, where possible. Using this list, synonyms for a given species are converted to the currently used name.



The master taxonomy for plants is the USDA PLANTS Database (USDA, NRCS, 2022; <https://plants.usda.gov>), and the portions of this database included in the NEON plant master taxonomy list are those pertaining to native and naturalized plants present within the NEON sampling area. A sublist for each NEON domain includes those species with ranges that overlap the domain and nativity designations—introduced or native—in that part of the range. If a species is reported at a location outside of its known range, and the record proves reliable, the master taxonomy list is updated to reflect the distribution change. For more details on plant taxonomic handling, see Barnett et al. (2019). For more on the NEON plant master taxonomy list, see NEON.DOC.014042 (<https://data.neonscience.org/api/v0/documents/NEON.DOC.014042vK>).

### *Data wrangling decisions*

In the plant presence and percent cover NEON data product (DP1.10058.001), sampling at the 1 × 1 m scale also includes observations of abiotic and nontarget species ground cover (i.e., soil, water, and downed wood), so we removed records with `divDataType` as “otherVariables.” We also removed records whose `targetTaxaPresent` is N (i.e., a nontarget species). Additionally, for all spatial resolutions (i.e., 1-, 10-, and 100-m<sup>2</sup> data), any record lacking information critical for combining data within a plot and for a given sampling bout (i.e., `plotID`, `subplotID`, `boutNumber`, `endDate`, or `taxonID`) was dropped from the dataset. Furthermore, records without a definitive genus- or species-level `taxonID` (i.e., those representing unidentified morphospecies) were not included. To combine data from different spatial resolutions into one data frame, we created a pivot column entitled `sample_area_m2` (with possible values of 1, 10, and 100). Because of the nested sampling design of the plant data, to capture all records within a subplot at the 100-m<sup>2</sup> scale, we incorporated all data from both the 1- and 10-m<sup>2</sup> scales for that subplot. Similarly, to obtain all records within a plot at the 400-m<sup>2</sup> scale, we included all data from that plot. Species abundance information was only recorded as area coverage within 1 by 1 m subplots; however, users may use the frequency of a species across subplots within a plot or plots within a site as a proxy of its abundance if needed.

## Ticks and tick pathogens

### *NEON sampling design*

Tick sampling occurs in six distributed plots at each site, which are randomly chosen in proportion to NLCD land cover class (LeVan, Thibault, et al., 2019). Ticks are sampled by walking the perimeter of a 40 × 40 m plot using a

1 × 1 m drag cloth. Ideally, 160 m is sampled (the shortest straight line distance between corners), but the cloth can be dragged around obstacles if a straight line is not possible. The acceptable total sampling area is between 80 and 180 m per plot. The cloth can also be flagged over vegetation when the cloth cannot be dragged across it. Ticks are collected from the cloth and technicians’ clothing at appropriate intervals, depending on vegetation density, and at every corner of the plot. Specimens are immediately transferred to a vial containing 95% ethanol.

Onset and offset of tick sampling coincide with phenological milestones at each site, beginning within 2 weeks of the onset of green-up and ending within 2 weeks of vegetation senescence (LeVan, Thibault, et al., 2019). Sampling bouts are only initiated if the high temperature on the two consecutive days prior to planned sampling was >0°C. Early-season sampling is conducted on a low-intensity schedule, with one sampling bout every 6 weeks. When more than five ticks of any life stage have been collected within the last calendar year at a site, sampling switches to a high-intensity schedule at the site—with one bout every 3 weeks. A site remains on the high-intensity schedule until fewer than five ticks are collected within a calendar year; then, sampling reverts back to the low-intensity schedule.

Ticks are sent to an external facility for identification to species, life stage, and sex (LeVan, Thibault, et al., 2019). A subset of nymphal ticks are additionally sent to a pathogen testing facility. *Ixodes* species are tested for *Anaplasma phagocytophilum*, *Babesia microti*, *Borrelia burgdorferi* sensu lato, *Borrelia miyamotoi*, *Borrelia mayonii*, other *Borrelia* species (*Borrelia* sp.), and a *Ehrlichia muris*-like agent (Pritt et al., 2017). Non-*Ixodes* species are tested for *A. phagocytophilum*, *Borrelia lonestari* (and other undefined *Borrelia* species), *Ehrlichia chaffeensis*, *Ehrlichia ewingii*, *Francisella tularensis*, and *Rickettsia rickettsii*. Additional information about tick pathogen testing can be found in the Tick Pathogen Testing SOP ([https://data.neonscience.org/api/v0/documents/UMASS\\_LMZ\\_tickPathogens\\_SOP\\_20160829](https://data.neonscience.org/api/v0/documents/UMASS_LMZ_tickPathogens_SOP_20160829)) for the NEON Tick-borne Pathogen Status data product.

### *Data wrangling decisions*

The tick NEON data product (DP1.10093.001) consists of two dataframes: “`tck_taxonomyProcessed`,” hereafter referred to as “taxonomy data”; and “`tck_fielddata`,” hereafter referred to as “field data.” Users should be aware of some issues related to taxonomic ID. Counts assigned to higher taxonomic levels (e.g., at the order-level Ixodida; IXOSP2) are not the sum of lower levels; rather, they represent the counts of individuals that could not reliably be assigned to a lower taxonomic unit. Samples that were

not identified in the laboratory were assigned to the highest taxonomic level (order Ixodida; IXOSP2). However, users could make an informed decision to assign these ticks to the most probable group if a subset of individuals from the same sample were assigned to a lower taxonomy.

To clean the tick data, we first removed surveys and samples not meeting quality standards. In the taxonomy data, we removed samples where sample condition was not listed as “OK” (<1% of records). In the field data, we removed records where samples were not collected due to logistical concerns (10%). We then combined male and female counts in the taxonomy table into one “adult” class. The taxonomy table was reformatted so that every row contained a sampleID and counts for each species life stages were separate columns (i.e., “wide format”). Next, we joined the field data to the taxonomy data, using the sample ID to link the two tables. When joining, we retained field records where no ticks were found in the field, and thus, there were no associated taxonomy data. In drags where ticks were not found, counts were given zeros. All counts were standardized by area sampled.

Prior to 2019, both field surveyors and laboratory taxonomists enumerated each tick life stage; consequently, in the joined dataset there were two sets of counts (“field counts” and “laboratory counts”). However, starting in 2019, counts were performed by taxonomists rather than field surveyors. Field surveys conducted after 2019 no longer have field counts. Users of tick abundance data should be aware that this change in protocol has several implications for data wrangling and for analysis. First, after 2019, tick counts are no longer published at the same time as field survey data. Subsequently, some field records from the most recent years have tick presence recorded (targetTaxaPresent = “Y”), but do not yet have associated counts or taxonomic information and so the counts are still listed as NA. Users should be aware that counts of zero are therefore published earlier than positive counts. We strongly urge users to filter data to those years where there are no counts pending.

The second major issue is that in years where both field counts and laboratory counts were available, they did not always agree (8% of records). In cases of disagreement, we generally used laboratory counts in the final abundance data, because this is the source of all tick count data after 2019 and because life stage identification was more accurate. However, there were a few exceptions where we used field count data. In some cases, only a subsample of a certain life stage was counted in the laboratory, which resulted in higher field counts than laboratory counts. In this case, we assigned the additional unidentified individuals (e.g., the difference between the field and laboratory counts) to the

order level (IXOSP2). If quality notes from NEON described ticks being lost in transit, we also added the additional lost individuals to the order level. There were some cases (<1%) where the field counts were greater than laboratory counts by more than 20% and where the explanation was not obvious; we removed these records. We note that the majority of samples (~85%) had no discrepancies between the laboratory or field; therefore, this process could be ignored by users whose analyses are not sensitive to exact counts.

The tick pathogen NEON data product (DPI.10092.001) consists of two dataframes: `tck_pathogen`, hereafter referred to as “pathogen data”; and `tck_pathogenqa`, hereafter referred to as “quality data.” First, we removed any samples that had flagged quality checks from the quality data and removed any samples that did not have a positive DNA quality check from the pathogen data. Although the original online protocol aimed to test 130 ticks per site per year from multiple tick species, the final sampling decision was to extensively sample IXOSCA, AMBAME, and AMBSP species only because IXOPAC and *Dermacentor* nymph frequencies were too rare to generate meaningful pathogen data. *Borrelia burgdorferi* and *B. burgdorferi sensu lato* tests were merged, since the former was an incomplete pathogen name and refers to *B. burgdorferi sensu lato* as opposed to *sensu stricto* (Rudenko et al., 2011). Tick pathogen data are presented as positivity rate calculated as the number of positive tests per number of tests conducted for a given pathogen on ticks collected during a given sampling event.

## Aquatic organisms

### Aquatic macroinvertebrates

#### *NEON sampling design*

Aquatic macroinvertebrate sampling occurs three times/year at wadeable stream, river, and lake sites from spring through fall. The timing of sampling is site-specific and based on historical hydrological, meteorological, and phenological data including dates of known ice cover, growing degree days, and green-up and brown-down (Cawley et al., 2016). Samplers vary by habitat and include Surber, Hess, hand corer, modified kicknet, D-frame sweep, and petite Ponar samplers (Parker, 2019). Stream sampling occurs throughout the 1-km permitted reach in wadeable areas of the two dominant habitat types. Lake sampling occurs with a petite Ponar near buoy, inlet and outlet sensors, and D-frame sweeps in wadeable littoral zones. Riverine sample collections in deep waters or near instrument buoys are made with a



petite Ponar, and in littoral areas are made with a D-frame sweep or large woody debris sampler. In the field, samples are preserved in pure ethanol, and later in the domain support facility, glycerol is added to prevent the samples from becoming brittle. Samples are shipped from the domain facility to a taxonomy laboratory for sorting and identification to the lowest possible taxon (e.g., genus or species), and counts of each taxon per size are made to the nearest millimeter.

#### *Data wrangling decisions*

Aquatic macroinvertebrate data contained in the NEON data product DP1.20120.001 are subsampled and identified to the lowest practical taxonomic level, typically genus, by expert taxonomists in the `inv_taxonomyProcessed` table, measured to the nearest millimeter size class, and counted. Taxonomic naming has been standardized in the `inv_taxonomyProcessed` file, according to NEON's master taxonomy (<https://data.neonscience.org/taxonomic-lists>), removing any synonyms. We calculated macroinvertebrate density by dividing `estimatedTotalCount` (which includes the corrections for subsampling in the taxonomy laboratory) by `benthicArea` from the `inv_fieldData` table to return count per square meter of stream, lake, or river bottom (Chesney et al., 2021).

## Microalgae (periphyton and phytoplankton)

#### *NEON sampling design*

NEON collects periphyton samples from natural surface substrata (i.e., cobble, silt, woody debris) over a 1-km reach in streams and rivers, and in the littoral zone of lakes. Various collection methods and sampler types are used, depending on substrate (Parker, 2020). In lakes and rivers, periphyton are also collected from the most dominant substratum type in three areas within the littoral (i.e., shoreline) zone. Prior to 2019, littoral zone periphyton sampling occurred in five areas.

NEON collects three phytoplankton samples per sampling date using Kemmerer or Van Dorn samplers. In rivers, samples are collected near the sensor buoy and at two other deep-water points in the main channel. For lakes, phytoplankton are collected near the central sensor buoy and at two littoral sensors. Where lakes and rivers are stratified, each phytoplankton sample is a composite from one surface sample, one sample from the metalimnion (i.e., middle layer), and one sample from the bottom of the euphotic zone. For nonstratified lakes and nonwadeable streams, each phytoplankton sample is a composite from one surface sample, one sample just above the bottom of the euphotic zone, and one mid-euphotic zone sample, if the euphotic zone is >5 m deep.

All microalgal sampling occurs three times per year (i.e., spring, summer, and fall bouts) in the same sampling bouts as aquatic macroinvertebrates and zooplankton. In wadeable streams, which have variable habitats (e.g., riffles, runs, pools, and step pools), three periphyton samples are collected per bout in the dominant habitat type (five samples collected prior to 2019) and three per bout in the second most dominant habitat type. No two samples are collected from the sample habitat unit (i.e., the same riffle).

Samples are processed at the domain support facility and separated into subsamples for taxonomic analysis or for biomass measurements. Aliquots shipped to an external facility for taxonomic determination are preserved in glutaraldehyde or Lugol's iodine (before 2021). Aliquots for biomass measurements are filtered onto glass-fiber filters and processed for ash-free dry mass (AFDM).

#### *Data wrangling decisions*

The periphyton, seston, and phytoplankton NEON data product (DP1.20166.001) contains three dataframes for algae containing information on algae taxonomic identification, biomass, and related field data, which are hereafter referred to as `alg_tax_long`, `alg_biomass`, and `alg_field_data`. Algae within samples are identified to the lowest possible taxonomic resolution, usually species, by contracting laboratory taxonomists. Some specimens can only be identified to the genus or even class level, depending on the condition of the specimen. Ten percent of all samples are checked by a second taxonomist and are noted in the `qcTaxonomyStatus`. Taxonomic naming has been standardized in the `alg_tax_long` files, according to NEON's master taxonomy, removing nomenclatural synonyms. Abundance and cell/colony counts are determined for each taxon of each sample with counts of cells or colonies that are either corrected for sample volume or not (as indicated by `algalParameterUnit = "cellspersBottle"`).

We corrected sample units of `cellspersBottle` to density (Parker & Vance, 2020). First, we summed the preservative volume and the laboratory's recorded sample volume for each sample (from the `alg_biomass` file) and combined that with the `alg_tax_long` file using `sampleID` as a common identifier. Where samples in the `alg_tax_long` file were missing data in the `perBottleSampleVolume` field (measured after receiving samples at the external laboratory), we estimated the sample volume using NEON domain laboratory sample volumes (measured prior to shipping samples to the external laboratory). With this updated file, we combined it with `alg_field_data` to have the related field conditions, including benthic area sampled for each sample. `parentSampleID` was used for `alg_field_data` to join to the `alg_biomass` file's `sampleID` as `alg_field_data` only has `parentSampleID`. We then calculated cells per milliliter for the uncorrected taxon of each

sample, dividing `algalParameterValue` by the updated sample volume. Benthic sample results are expressed in terms of area (i.e., multiplied by the field sample volume and divided by benthic area sampled), in square meters. The final abundance units are either cells per milliliter (phytoplankton and seston samples) or cells per square meter for benthic samples.

The `sampleIDs` are child records of each `parentSampleID` that will be collected as long as sampling is not impeded (i.e., ice-covered or dry). In the `alg_biomass` file, there should be only a single entry for each `parentSampleID`, `sampleID`, and `analysisType`. Most often, there were two `sampleIDs` per `parentSampleID` with one for AFDM and taxonomy (analysis types). For the creation of the observation table with standardized counts, we used only records from the `alg_biomass` file with the `analysisType` of taxonomy. In `alg_tax_long`, there are multiple entries for each `sampleID` for each taxon by `scientificName` and `algalParameter`.

## Fish

### *NEON sampling design*

Fish sampling is carried out across 19 of the NEON ecoclimatic domains, occurring in a total of 23 lotic (stream) and 5 lentic (lake) sites. In lotic sites, up to 10 nonoverlapping reaches, each 70–130 m long, are designated within a 1-km section of stream (Jensen et al., 2019a). These include three constantly sampled “fixed” reaches, which encompass all representative habitats found within the 1-km stretch, and seven “random” reaches that are sampled on a rotating schedule. In lentic sites, 10 pie-shaped segments are established, with each segment ranging from the riparian zone into the lake center, therefore effectively capturing both nearshore and offshore habitats (Jensen et al., 2019b). Three of the 10 segments are fixed and are surveyed twice a year, and the remaining segments are random and are sampled rotationally. The spatial layouts of these sites are designed to capture spatial and temporal heterogeneity in the aquatic habitats.

Lotic sampling occurs at three fixed and three random reaches per sampling bout, and there are two bouts per year—one in spring and one in fall. During each bout, the fixed reaches are sampled via a three-pass electrofishing depletion approach (Moulton II et al., 2002; Peck et al., 2006), while the random reaches being sampled are done so with a single-pass depletion approach. Which random reaches are surveyed depends on the year, with three of the random reaches sampled every other year. All sampling occurs during daylight hours, with each sampling bout completed within 5 days and with a minimum 2-week gap in between two successive

sampling bouts. The initial sampling date is determined using site-specific historical data on ice melting, water temperature (or accumulated degree days), and riparian peak greenness.

The lentic sampling design is similar to that discussed above, with fixed segments being sampled twice per year and random segments sampled twice per year on a rotational basis (i.e., each random segment is not sampled every year). Lentic sampling is conducted using three gear types, with backpack electrofishing and mini-fyke nets near the shoreline and gill nets in deeper waters. Backpack electrofishing is done on a 4 × 25 m reach near the shoreline via a three-pass (for fixed segments) or single-pass (for random segments) electrofishing depletion approach (Moulton II et al., 2002, Peck et al., 2006). All three passes in a fixed sampling segment are completed on the same night, with ≤30 min between successive passes. Electro-fishing begins within 30 min of sunset and ceases within 30 min of sunrise, with a maximum of five passes per sampling bout. A single gill net is also deployed within all segments being sampled, both fixed and random, for 1–2 h in either the morning or early afternoon. Finally, a fyke (Baker et al., 1997) or mini-fyke net is deployed at each fixed or random segments, respectively. Fyke nets are positioned before sunset and recovered after sunrise on the following day. Precise start and end times for electrofishing and net deployments are documented by NEON technicians at the time of sampling.

In all surveys, captured fish are identified to the lowest practical taxonomic level, and morphometrics (i.e., body mass and body length) are recorded for 50 individuals of each taxon before releasing. Relative abundance for each fish taxon is also recorded by direct enumeration (up to first 50 individuals) or estimation by bulk counts (>50 individuals, i.e., by placing fish of a given taxon into a dip net [i.e., net scoop], counting the total number of specimens in the dip net, and then multiplying the total number of scoops of captured fish by the counts from the first scoop).

### *Data wrangling decisions*

Fish sampled via both electrofishing and trapping are identified at variable taxonomic resolutions (as fine as subspecies level) in the field. Most identifications are made to the species or genus level by a single field technician for a given bout per site. Sampled fish are identified, measured, weighed, and then released back to the site of capture. If field technicians are unable to identify to the species level, such specimens are identified to the finest possible taxonomic resolution or assigned a morphospecies with a coarse-resolution identification. The standard sources consulted for identification and a qualifier for identification validity are also documented in the `fish_perFish` table. The column

bulkFishCount of the fsh\_bulkCount table records relative abundance for each species or the alternative next possible taxon level (specified in the column scientificName).

Fish data (taxonomic identification and relative abundance) are recorded per each sampling reach in streams or per segment in lakes in each bout and documented in the fsh\_perFsh table (Monahan et al., 2020). The column eventID uniquely identifies the sampling date of the year, the specific site within the domain, a reach/segment identifier, the pass number (i.e., number of electrofishing passes or number of net deployment efforts), and the survey method. The eventID column helps tie all fish data with stream reach/lake segment data or environmental data (i.e., water quality data) and sampling effort data (e.g., electrofishing and net set time). A reachID column provided in the fsh\_perPass table uniquely identifies surveys done per stream reach or lake segment. The reachID is nested within the eventID as well. We used eventID as a nominal variable to uniquely identify different sampling events and to join different, stacked fish data files as described below.

The fish NEON data product (DP1.20107.001) consists of fsh\_perPass, fsh\_fieldData, fsh\_bulkCount, fsh\_perFish, and the complete taxon table for fish, for both stream and lake sites. To join all reach-scale data, we first joined the fsh\_perPass with fsh\_fieldData, and eliminated all bouts where sampling was untenable. Subsequently, we joined the reach-scale table with fsh\_perFsh to add individual fish counts and fish measurements. Then, to add bulk counts, we joined the reach-scale table with fsh\_bulkCount datasets, and subsequently added taxonRank, which included the taxonomic resolution in the bulk-processed table. Afterward, both individual-level and bulk-processed datasets were appended into a single table. To include samples where no fish were captured, we filtered the fsh\_perPass table retaining records where target taxa (fish) were absent, joined it with fsh\_fieldData, and finally merged it with the table that contained both bulk-processed and individual-level data. For each finer-resolution taxon in the individual-level dataset, we considered the relative abundance as one since each row represented a single individual fish. Whenever possible, we substituted missing data by cross-referencing other data columns, omitted completely redundant data columns, and retained records with genus- and species-level taxonomic resolution. For the appended dataset, we also calculated the relative abundance for each species per sampling reach or segment at a given site. To calculate species-specific catch per unit effort (CPUE), we normalized the relative abundance by either average electrofishing time (i.e., efTime, efTime2) or trap deployment time (i.e., the difference between netEndTime and netSetTime). For trap data, we assumed size of the traps used, water depths, number of netters used, and the reach lengths (a significant

proportion of bouts had reach lengths missing) to be comparable across different sampling reaches and segments.

## Zooplankton

### *NEON sampling design*

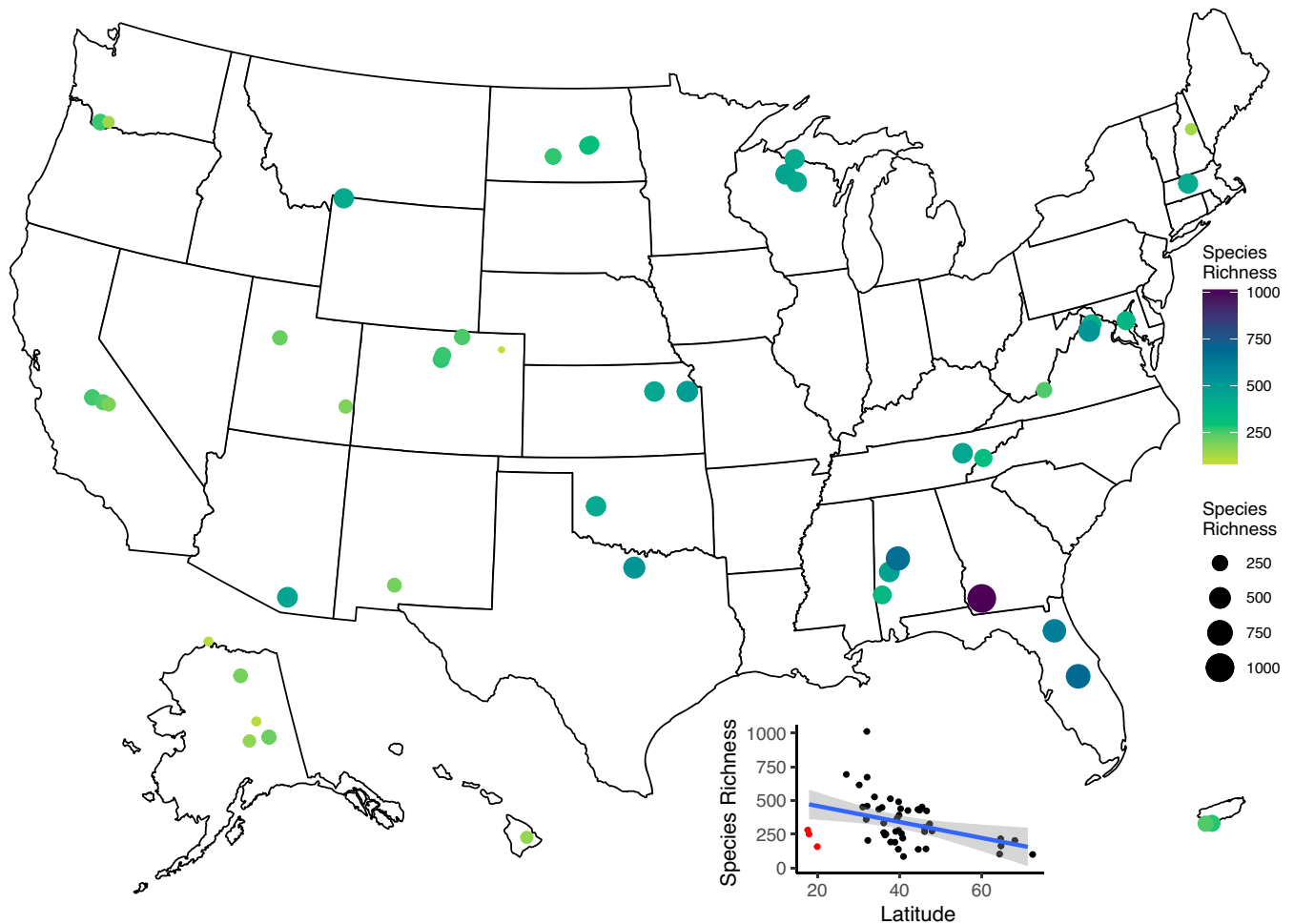
Zooplankton samples are collected at seven NEON lake sites across four domains. Zooplankton samples are collected at the buoy sensor set (deepest location in the lake) and at the two nearshore sensor sets using a vertical tow net for locations deeper than 4 m and a Schindler trap for locations shallower than 4 m (Parker & Roehm, 2019). This results in three samples collected per sampling day. Samples are preserved with ethanol in the field and shipped from the domain facility to a taxonomy laboratory for sorting and identification to the lowest possible taxon (e.g., genus or species), and counts of each taxon per size are made to the nearest millimeter.

### *Data wrangling decisions*

The NEON zooplankton data product (DP1.20219.001) consists of dataframes for taxonomic identification and related field data (Parker & Scott, 2020). Zooplankton in NEON samples are identified at contracting laboratories to the lowest possible taxonomic resolution, usually genus; however, some specimens can only be identified to the family (or even class) level, depending on the condition of the specimen. Ten percent of all samples are checked by two taxonomists and are noted in the qcTaxonomyStatus column. The taxonomic naming has been standardized in the zoo\_taxonomyProcessed table, according to NEON's master taxonomy, removing any synonyms. Density was calculated using adjCountPerBottle and towsTrapsVolume to correct count data to "count per liter."

## RESULTS (OR HOW TO GET AND USE STANDARDIZED NEON ORGANISMAL DATA)

All cleaned and standardized datasets can be obtained from the R package neonDivData and from the EDI data repository (<https://doi.org/10.6073/pasta/c28dd4f6e7989003505ea02e9a92afb>). Note that neonDivData included both stable and provisional data released by NEON, while the data repository in EDI only included stable datasets. If users want to change some of the decisions to wrangle the data differently, they can find the code in the R package ecomDP and modify them for their own purposes. If this standardized version of NEON data was used, users should cite this paper along with the citations provided by



**FIGURE 2** Plant species richness mapped across NEON terrestrial sites. The inset scatterplot shows latitude on the *x*-axis and species richness on the *y*-axis, with red points representing sites in Puerto Rico and Hawaii.

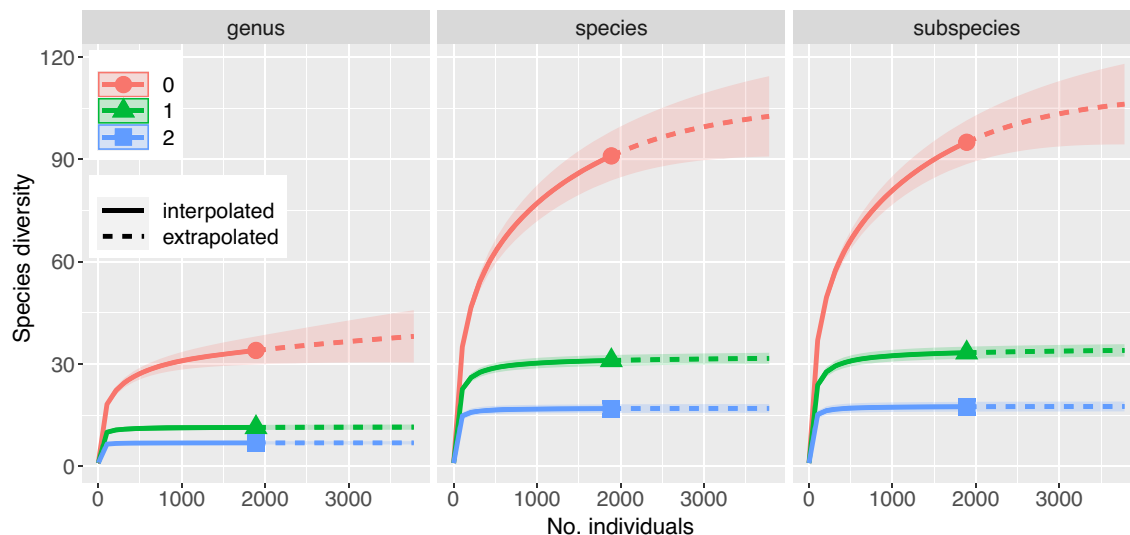
NEON for each taxonomic group. Such citations can be found in the URLs presented in Table 1.

The data package `neonDivData` can be installed from GitHub. Installation instructions can be found on the GitHub webpage (<https://github.com/daijiang/neonDivData>). Table 2 shows a brief summary of all data objects. To get data for a specific taxonomic group, we can just call the objects in the R object column in Table 2. Such data products include cleaned (and standardized if needed) occurrence data for the taxonomic groups covered and are equivalent to the “observation” table of the `ecocomDP` data format. If environmental information was provided by NEON for some taxonomic groups, they are also included in these data objects. Information such as latitude, longitude, and elevation for all taxonomic groups was saved in the `neon_location` object of the R package, which is equivalent to the “sampling\_location” table of the `ecocomDP` data format. Information about species scientific names of all taxonomic groups was saved in the `neon_taxa` object, which is equivalent to the “taxon” table of the `ecocomDP` data format.

To demonstrate the use of data packages, we used `data_plant` to quickly visualize the distribution of species richness of plants across all NEON sites (Figure 2). To show how easy it is to get site-level species richness, we presented the code used to generate the data for Figure 2 as supporting information.

Figure 2 shows the utility of the data package for exploring macroecological patterns. One of the most well-known and studied macroecological patterns is the latitudinal biodiversity gradient, wherein sites are more species-rich at lower latitudes relative to higher latitudes; temperature, biotic interactions, and historical biogeography are potential reasons underlying these patterns (Fischer, 1960; Hillebrand, 2004). Herbaceous plants of NEON generally follow this pattern. The latitudinal pattern for NEON small mammals is similar and is best explained by increased niche space and declining similarity in body size among species in lower latitudes, rather than a direct effect of temperature (Read et al., 2018).

In addition to allowing for quick exploration of macroecological patterns of richness at NEON sites, the



**FIGURE 3** Rarefaction of beetle abundance data from collections made at the Oak Ridge National Laboratory NEON site from 2014 to 2020 generated using the beetle data package presented in this paper and the iNEXT package in R (Hsieh et al., 2016) based on different levels of taxonomic resolution (i.e., genus, species, and subspecies). Different colors indicate Jost's indices with differing values of  $q$  (Jost, 2006).

data packages presented in this paper enable investigation of the effects of taxonomic resolution on diversity indices since taxonomic information is preserved for observations under family level for all groups. The degree of taxonomic resolution varies for NEON taxa depending on the diversity of the group and the level of taxonomic expertise needed to identify an organism to the species level, with more diverse groups presenting a greater challenge. Beetles are one of the most diverse groups of organisms on Earth and wide-ranging geographically, making them ideal bioindicators of environmental change (Rainio & Niemelä, 2003). To illustrate how the use of the beetle data package presented in this paper enables NEON data users to easily explore the effects of taxonomic resolution on community-level taxonomic diversity metrics, we calculated Jost's diversity indices (Jost, 2006) for beetles at the Oak Ridge National Laboratory (ORNL) NEON site for data subset at the genus, species, and subspecies level. To quantify biodiversity, we used Jost's indices, which are essentially Hill numbers that vary in how abundance is weighted with a parameter  $q$ . Higher values of  $q$  give lower weights to low-abundance species, with  $q = 0$  being equivalent to species richness and  $q = 1$  representing the effective number of species given by the Shannon entropy. These indices are plotted as rarefaction curves, which assess the sampling efficacy. When rarefaction curves asymptote, they suggest that additional sampling will not capture additional taxa. Statistical methods presented by Chao et al. (2014) provide estimates of sampling efficacy beyond the observed data (i.e., extrapolated values shown by dashed lines in

Figure 3). For the ORNL beetle data, Jost's indices calculated with higher values of  $q$  (i.e.,  $q > 0$ ) indicated sampling has reached an asymptote in terms of capturing diversity regardless of taxonomic resolution (i.e., genus, species, and subspecies). However, rarefaction curves for  $q = 0$ , which is equivalent to species richness, do not asymptote, even with extrapolation. These plots suggest that if a researcher is interested in low-abundance, rare species, then the NEON beetle data stream at ORNL may need to mature with additional sample collections over time before confident inferences may be made, especially below the taxonomic resolution of the genus.

## DISCUSSION (OR HOW TO MAINTAIN AND UPDATE STANDARDIZED NEON ORGANISMAL DATA)

NEON organismal data hold enormous potential to understand biodiversity change across space and time (Balch et al., 2019; Jones et al., 2021). Multiple biodiversity research and education programs have used NEON data even before NEON became fully operational in May 2019 (e.g., Farrell & Carey, 2018; Read et al., 2018). With the expected long-term investment to maintain NEON over the next 30 years, NEON organismal data will be an invaluable tool for understanding and tracking biodiversity change. NEON data are unique relative to data collected by other similar networks (e.g., LTER, CZO) because observation collection protocols are standardized



across sites, enabling researchers to address macroscale questions in environmental science without having to synthesize disparate datasets that differ in collection methods (Jones et al., 2021). The data package presented in this paper holds great potential in making NEON data easier to use and more comparable across studies. Whereas the data collection protocols implemented by NEON staff are standardized, the decisions NEON data users make in wrangling their data after downloading NEON's open data will not necessarily be similar unless the user community adopts a community data standard, such as the *ecocomDP* data model. Adopting such a data model early on in the life of the observatory will ensure that the results of studies using NEON data will be comparable and thus easier to synthesize. By providing a standardized and easy-to-use data package of NEON organismal data, our effort here will significantly lower the barriers to use the NEON organismal data for biodiversity research by many current and future researchers and will ensure that studies using NEON organismal data are comparable.

All code for the Data Wrangling Decisions are available within the R package *ecocomDP* (<https://github.com/EDIorg/ecocomDP>). Users can modify the code if they need to make different decisions during the data wrangling process and update our workflows in our code by submitting a pull request to our GitHub repository. If researchers wish to generate their own derived organismal datasets from NEON data with slightly different decisions than the ones outlined in this paper, we recommend that they use the *ecocomDP* framework, contribute their workflow to the *ecocomDP* R package, upload the data to the EDI repository, and cite their data with the discoverable DOI given to them by EDI. Note that the *ecocomDP* data model was intended for community ecology analyses and may not be well suited for population-level analyses. In a similar vein, researchers should ensure that they have considered sample size issues before fitting any models with these data. See Barnett (2019) for a review of the NEON organismal sampling design that contains important insights related to sample size issues.

Because *ecocomDP* is an R package to access and format datasets following the *ecocomDP* format, we developed an R data package *neonDivData* to host and distribute the standardized NEON organismal data derived from *ecocomDP*. A separate dedicated data package has several advantages. First, it is easier and ready to use and saves time for users to run the code in *ecocomDP* to download and standardize NEON data products. Second, it is also easy to update the data package when new raw data products are uploaded by NEON to their data portal, and the updating process does not require any

change in the *ecocomDP* package. This is ideal because *ecocomDP* provides harmonized data from other sources besides NEON. Third, the GitHub repository page of *neonDivData* can serve as a discussion forum for researchers regarding the NEON data products without competing for attention in the *ecocomDP* GitHub repository page. By opening issues on the GitHub repository, users can discuss and contribute to improve our workflow of standardizing NEON data products. Users can also discuss whether there are other data models that the NEON user community should adopt at the inception of the observatory. As the observatory moves forward, this is an important discussion for the NEON user community and NEON technical working groups to promote the synthesis of NEON data with data from other efforts (e.g., LTER, CZO, AmeriFlux, the International LTER, National Phenology Network, and Long Term Agricultural Research Network). Note that the standardized datasets that are stable (defined by NEON as stable release) were archived at EDI and some of the above advantages also apply to the data repository at EDI.

The derived data products presented here collectively represent hundreds of hours of work by members of our team—a group that met at the NEON Science Summit in 2019 in Boulder, Colorado, and consists of researchers and NEON science staff. Just as it is helpful when working with a dataset to either have collected the data or be in close correspondence with the person who collected the data, final processing decisions were greatly informed by conversations with NEON science staff and the NEON user community. Future opportunities that encourage collaborations between NEON science staff and the NEON user community will be essential to achieve the full potential of the observatory data.

## CONCLUSION

Macrosystems ecology (*sensu* Heffernan et al., 2014) is at the start of an exciting new chapter with the decades-long awaited buildout of NEON completed and standardized data streams from all sites in the observatory becoming publicly available online. As the research community embarks on discovering new scientific insights from NEON data, it is important that we make our analyses and all derived data as reproducible as possible to ensure that connections across studies are possible. Harmonized datasets will help in this endeavor because they naturally promote the collection of provenance as data are collated into derived products (O'Brien et al., 2021; Reichman et al., 2011). Harmonized data also make synthesis easier because efforts to clean and format data leading up to analyses do not have to be repeatedly performed by



individual researchers (O'Brien et al., 2021). The data standardizing processes and derived data package presented here illustrate a potential path forward in achieving a reproducible framework for data derived from NEON organismal data for ecological analyses. This derived data package also highlights the value of collaboration between the NEON user community and NEON staff for advancing NEON-enabled science. Finally, the extension of the ecomDP harmonized data design pattern to data from other ecological research and observatory networks (e.g., the Brazilian Network of Networks; de Oliveira Roque et al., 2018) and South African Environment Observation Network (Van Jaarsveld et al., 2007) has the potential to enable community ecologists to better synthesize data from across the globe.

## ACKNOWLEDGMENTS

This work is a result of participating in the first NEON Science Summit in 2019 and an internship program through the St. Edward's Institute for Interdisciplinary Science (i4) funded through a National Science Foundation (NSF) award under grant number 1832282. The authors acknowledge support from the NSF Award 1906144 to attend the 2019 NEON Science Summit. Additionally, the authors acknowledge support from the NSF DEB 1926568 to Sydne Record, NSF DEB 1926567 to Phoebe Zarnetske, NSF DEB 1926598 to Marta A. Jarzyna, and NSF DEB 1926341 to Jalene M. LaMontagne. Comments from NEON staff (Katie LeVan, Dylan Mpnahan, Sata Paull, Dave Barnett, and Sam Simkin), Margaret O'Brien, and Tad Dallas greatly improved this work. The NEON is a program sponsored by the NSF and operated under a cooperative agreement by Battelle Memorial Institute. This material is based in part upon work supported by the NSF through the NEON Program.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

No data were collected for this study. All original data were collected by NEON and are publicly available at NEON's data portal. We standardized such data and provided them as an R package available from Zenodo: <https://doi.org/10.5281/zenodo.6419751>. Other data (Li et al., 2022) are available from the Environmental Data Initiative (EDI) Data Portal: <https://doi.org/10.6073/pasta/c28dd4f6e7989003505ea02e9a92afbf>.

## ORCID

Daijiang Li  <https://orcid.org/0000-0002-0925-3421>  
 Sydne Record  <https://orcid.org/0000-0001-7293-2155>  
 Eric R. Sokol  <https://orcid.org/0000-0001-5923-0917>

Y. Anny Chung  <https://orcid.org/0000-0001-5207-2872>  
 Matthew R. Helmus  <https://orcid.org/0000-0003-3977-0507>  
 Marta A. Jarzyna  <https://orcid.org/0000-0002-6734-0566>  
 Michael G. Just  <https://orcid.org/0000-0003-2493-9269>  
 Jalene M. LaMontagne  <https://orcid.org/0000-0001-7713-8591>  
 Brett A. Melbourne  <https://orcid.org/0000-0002-8843-4131>  
 Wynne Moss  <https://orcid.org/0000-0002-2813-1710>  
 Kari E. A. Norman  <https://orcid.org/0000-0002-2029-2325>  
 Bijan Seyednasrollah  <https://orcid.org/0000-0002-5195-2074>  
 Colin Smith  <https://orcid.org/0000-0003-2261-9931>  
 Sarah Spaulding  <https://orcid.org/0000-0002-9787-7743>  
 Thilina D. Surasinghe  <https://orcid.org/0000-0001-8758-072X>  
 Sarah K. Thomsen  <https://orcid.org/0000-0001-5964-7536>  
 Phoebe L. Zarnetske  <https://orcid.org/0000-0001-6257-6951>

## REFERENCES

- Baker, J. R., D. V. Peck, and D. W. Sutton. 1997. *Environmental Monitoring and Assessment Program Surface Waters: Field Operations Manual for Lakes*. Washington, DC: US Environmental Protection Agency.
- Balch, J. K., R. Nagy, and B. S. Halpern. 2019. "NEON Is Seeding the Next Revolution in Ecology." *Frontiers in Ecology and the Environment* 18: 3.
- Barnett, D. 2019. *TOS Protocol and Procedure: DIV—Plant Diversity Sampling*. NEON.DOC.014042vK. Boulder, CO: NEON.
- Barnett, D. T., P. B. Adler, B. R. Chemel, P. A. Duffy, B. J. Enquist, J. B. Grace, S. Harrison, et al. 2019. "The Plant Diversity Sampling Design for the National Ecological Observatory Network." *Ecosphere* 10: e02603.
- Bechtold, W. A., and P. L. Patterson. 2005. *The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures*. Asheville, NC: USDA Forest Service, Southern Research Station.
- Beck, J., M. Böller, A. Erhardt, and W. Schwanghart. 2014. "Spatial Bias in the GBIF Database and Its Effect on Modeling Species' Geographic Distributions." *Ecological Informatics* 19: 10–5.
- Blowes, S. A., S. R. Supp, L. H. Antão, A. Bates, H. Bruelheide, J. M. Chase, F. Moyes, et al. 2019. "The Geography of Biodiversity Change in Marine and Terrestrial Assemblages." *Science* 366: 339–45.
- Brown, J. H., J. F. Gillooly, A. P. Allen, V. M. Savage, and G. B. West. 2004. "Toward a Metabolic Theory of Ecology." *Ecology* 85: 1771–89.
- Cawley, K. M., S. Parker, R. Utz, K. Goodman, C. Scott, M. Fitzgerald, J. Vance, B. Jensen, C. Bohall, and T. Baldwin. 2016. *NEON Aquatic Sampling Strategy*. NEON.DOC.001152vA. Boulder, CO: NEON.

- Chao, A., N. J. Gotelli, T. Hsieh, E. L. Sander, K. Ma, R. K. Colwell, and A. M. Ellison. 2014. "Rarefaction and Extrapolation with Hill Numbers: A Framework for Sampling and Estimation in Species Diversity Studies." *Ecological Monographs* 84: 45–67.
- Chesney, T., S. Parker, and C. Scott. 2021. *NEON User Guide to Aquatic Macroinvertebrate Collection (DPI.20120.001). Revision b*. Boulder, CO: NEON.
- Curtis, J. T. 1959. *The Vegetation of Wisconsin: An Ordination of Plant Communities*. Madison, WI: University of Wisconsin Press.
- Egli, L., K. E. LeVan, and T. T. Work. 2020. "Taxonomic Error Rates Affect Interpretations of a National-Scale Ground Beetle Monitoring Program at National Ecological Observatory Network." *Ecosphere* 11: e03035.
- Farley, S. S., A. Dawson, S. J. Goring, and J. W. Williams. 2018. "Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions." *BioScience* 68: 563–76.
- Farrell, K. J., and C. C. Carey. 2018. "Power, Pitfalls, and Potential for Integrating Computational Literacy into Undergraduate Ecology Courses." *Ecology and Evolution* 8: 7744–51.
- Fischer, A. G. 1960. "Latitudinal Variations in Organic Diversity." *Evolution* 14: 64–81.
- Geldmann, J., J. Heilmann-Clausen, T. E. Holm, I. Levinsky, B. Markussen, K. Olsen, C. Rahbek, and A. P. Tøttrup. 2016. "What Determines Spatial Bias in Citizen Science? Exploring Four Recording Schemes with Different Proficiency Requirements." *Diversity and Distributions* 22: 1139–49.
- Gurevitch, J., and L. V. Hedges. 1999. "Statistical Issues in Ecological Meta-Analyses." *Ecology* 80: 1142–9.
- Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter. 2013. "Big Data and the Future of Ecology." *Frontiers in Ecology and the Environment* 11: 156–62.
- Harte, J. 2011. *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*. Oxford: Oxford University Press.
- Heffernan, J. B., P. A. Soranno, M. J. Angilletta, Jr., L. B. Buckley, D. S. Gruner, T. H. Keitt, J. R. Kellner, et al. 2014. "Macro-systems Ecology: Understanding Ecological Patterns and Processes at Continental Scales." *Frontiers in Ecology and the Environment* 12: 5–14.
- Hillebrand, H. 2004. "On the Generality of the Latitudinal Diversity Gradient." *The American Naturalist* 163: 192–211.
- Hoekman, D., K. E. LeVan, C. Gibson, G. E. Ball, R. A. Browne, R. L. Davidson, T. L. Erwin, et al. 2017. "Design for Ground Beetle Abundance and Diversity Sampling within the National Ecological Observatory Network." *Ecosphere* 8: e01744.
- Hsieh, T., K. Ma, and A. Chao. 2016. "iNEXT: An R Package for Rarefaction and Extrapolation of Species Diversity (Hill Numbers)." *Methods in Ecology and Evolution* 7: 1451–6.
- Hubbell, S. P. 2001. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton, NJ: Princeton University Press.
- Hutchinson, G. E. 1959. "Homage to Santa Rosalia or Why Are there So Many Kinds of Animals?" *The American Naturalist* 93: 145–59.
- Jensen, B., S. Parker, and J. R. Fischer. 2019a. *AOS Protocol and Procedure: Fish Sampling in Wadeable Streams*. NEON.DOC.001295vF. Boulder, CO: NEON.
- Jensen, B., S. Parker, and J. R. Fischer. 2019b. *AOS Protocol and Procedure: Fish Sampling in Lakes*. NEON.DOC.001296vF. Boulder, CO: NEON.
- Jones, J., P. Groffman, J. Blair, F. Davis, H. Dugan, E. Euskirchen, S. Frey, et al. 2021. "Synergies among Environmental Science Research and Monitoring Networks: A Research Agenda." *Earth's Future* 9(3): e2020EF001631.
- Jost, L. 2006. "Entropy and Diversity." *Oikos* 113: 363–75.
- Keller, M., D. S. Schimel, W. W. Hargrove, and F. M. Hoffman. 2008. "A Continental Strategy for the National Ecological Observatory Network." *Frontiers in Ecology and the Environment* 6(5): 282–4.
- Koricheva, J., and J. Gurevitch. 2014. "Uses and Misuses of Meta-Analysis in Plant Ecology." *Journal of Ecology* 102: 828–44.
- LeVan, K. 2020a. *NEON User Guide to Ground Beetles Sampled from Pitfall Traps (DPI.10022.001). Version c*. Boulder, CO: NEON.
- LeVan, K. 2020b. *NEON User Guide to Mosquitoes Sampled from CO<sub>2</sub> Traps (DPI.10043.001) and Mosquito-Borne Pathogen Status (DPI.10041.001). Version c*. Boulder, CO: NEON.
- LeVan, K., S. Paull, K. Tsao, D. Hoekman, and Y. Springer. 2019. *TOS Protocol and Procedure: MOS—Mosquito Sampling*. NEON.DOC.014049vL. Boulder, CO: NEON.
- LeVan, K., N. Robinson, D. Hoekman, and K. Blevins. 2019. *TOS Protocol and Procedure: Ground Beetle Sampling*. NEON.DOC.014041vJ. Boulder, CO: NEON.
- LeVan, K., K. Thibault, K. Tsao, and Y. Springer. 2019. *TOS Protocol and Procedure: Tick and Tick-Borne Pathogen Sampling*. NEON.DOC.014045vK. Boulder, CO: NEON.
- Li, D., J. D. Olden, J. L. Lockwood, S. Record, M. L. McKinney, and B. Baiser. 2020. "Changes in Taxonomic and Phylogenetic Diversity in the Anthropocene." *Proceedings of the Royal Society B* 287: 20200777.
- Li, D., S. Record, E. Sokol, M. E. Bitters, M. Y. Chen, A. Y. Chung, M. R. Helmus, et al. 2022. "Standardized NEON Organismal Data (neonDivData) Ver 1." <https://doi.org/10.6073/pasta/c28dd4f6e7989003505ea02e9a92afbf>
- Linnaeus, C. 1758. *Systema Naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Editio decima, reformata, vol. 1, 824*. Stockholm: Laurentius Salvius.
- MacArthur, R. H., and E. O. Wilson. 1967. *The Theory of Island Biogeography*. Princeton, NJ: Princeton University Press.
- Martin, L. J., B. Blossey, and E. Ellis. 2012. "Mapping where Ecologists Work: Biases in the Global Distribution of Terrestrial Ecological Observations." *Frontiers in Ecology and the Environment* 10: 195–201.
- Midgley, G. F., and W. Thuiller. 2005. "Global Environmental Change and the Uncertain Fate of Biodiversity." *The New Phytologist* 167: 638–41.
- Monahan, D., B. Jensen, S. Parker, and C. Scott. 2020. *NEON User Guide to Fish Electrofishing, Gill Netting, and Fyke Netting Counts (DPI.20107.001). Revision b*. Boulder, CO: NEON.
- Moulton, S. R., II, J. G. Kennen, R. M. Goldstein, and J. A. Hambrook. 2002. *Revised Protocols for Sampling Algal, Invertebrate, and Fish Communities as Part of the National Water-Quality Assessment Program*. Reston, VA: US Geological Survey. Open-File Report 2002-150.
- Nakagawa, S., and E. S. Santos. 2012. "Methodological Issues and Advances in Biological Meta-Analysis." *Evolutionary Ecology* 26: 1253–74.
- O'Brien, M., C. A. Smith, E. R. Sokol, C. Gries, N. Lany, S. Record, and M. C. Castorani. 2021. "ecocomDP: A Flexible Data Design Pattern for Ecological Community Survey Data." *Ecological Informatics* 64: 101374.

- de Oliveira Roque, F., J. F. Menezes, T. Northfield, J. M. Ochoa-Quintero, M. J. Campbell, and W. F. Laurance. 2018. “Warning Signals of Biodiversity Collapse across Gradients of Tropical Forest Loss.” *Scientific Reports* 8: 1–7.
- Palumbo, I., R. A. Rose, R. M. Headley, J. Nackoney, A. Vodacek, and M. Wegmann. 2017. “Building Capacity in Remote Sensing for Conservation: Present and Future Challenges.” *Remote Sensing in Ecology and Conservation* 3: 21–9.
- Parker, S. 2019. *AOS Protocol and Procedure: INV—Aquatic Macroinvertebrate Sampling*. NEON.DOC.003046vE. Boulder, CO: NEON.
- Parker, S. 2020. *AOS Protocol and Procedure: ALG—Periphyton and Phytoplankton Sampling*. NEON.DOC.003045vE. Boulder, CO: NEON.
- Parker, S., and C. Roehm. 2019. *AOS Protocol and Procedure: ZOO—Zooplankton Sampling in Lakes*. NEON.DOC.001194. Boulder, CO: NEON.
- Parker, S., and C. Scott. 2020. *NEON User Guide to Aquatic Zooplankton Collection (DP1.20219.001)*. Revision b. Boulder, CO: NEON.
- Parker, S., and T. Vance. 2020. *NEON User Guide to Periphyton and Phytoplankton Collection (DP1.20166.001)*. Revision c. Boulder, CO: NEON.
- Pavlacky, D. C., Jr., P. M. Lukacs, J. A. Blakesley, R. C. Skorkowsky, D. S. Klute, B. A. Hahn, V. J. Dreitz, T. L. George, and D. J. Hanni. 2017. “A Statistically Rigorous Sampling Design to Integrate Avian Monitoring and Management within Bird Conservation Regions.” *PLoS One* 12: e0185924.
- Peck, D. V., A. T. Herlihy, B. H. Hill, R. M. Hughes, P. R. Kaufmann, D. J. Klemm, J. M. Lazorchak, et al. 2006. *Environmental Monitoring and Assessment Program—Surface Waters: Western Pilot Study Field Operations Manual for Wadeable Streams*. Washington, DC: US Environmental Protection Agency.
- Pricope, N. G., K. L. Mapes, and K. D. Woodward. 2019. “Remote Sensing of Human–Environment Interactions in Global Change Research: A Review of Advances, Challenges and Future Directions.” *Remote Sensing* 11: 2783.
- Pritt, B. S., M. E. Allerdice, L. M. Sloan, C. D. Paddock, U. G. Munderloh, Y. Rikihisa, T. Tajima, et al. 2017. “Proposal to Reclassify *Ehrlichia muris* as *Ehrlichia muris* Subsp. *muris* Subsp. nov. and Description of *Ehrlichia muris* Subsp. *eauclairensis* Subsp. nov., a Newly Recognized Tick-Borne Pathogen of Humans.” *International Journal of Systematic and Evolutionary Microbiology* 67: 2121.
- Qin, C., R. Bartelme, Y. A. Chung, D. Fairbanks, Y. Lin, D. Liptzin, C. Muscarella, et al. 2021. “From DNA Sequences to Microbial Ecology: Wrangling NEON Soil Microbe Data with the neonMicrobe r Package.” *Ecosphere* 12: e03842.
- Rainio, J., and J. Niemelä. 2003. “Ground Beetles (Coleoptera: Carabidae) as Bioindicators.” *Biodiversity and Conservation* 12: 487–506.
- Ralph, C. J. 1993. *Handbook of Field Methods for Monitoring Landbirds*. Albany, CA: Pacific Southwest Research Station.
- Read, Q. D., J. M. Grady, P. L. Zarnetske, S. Record, B. Baiser, J. Belmaker, M.-N. Tuanmu, A. Strecker, L. Beaudrot, and K. M. Thibault. 2018. “Among-Species Overlap in Rodent Body Size Distributions Predicts Species Richness along a Temperature Gradient.” *Ecography* 41: 1718–27.
- Record, S., N. M. Voelker, P. L. Zarnetske, N. I. Wisnoski, J. D. Tonkin, C. Swan, L. Marazzi, et al. 2020. “Novel Insights to Be Gained from Applying Metacommunity Theory to Long-Term, Spatially Replicated Biodiversity Data.” *Frontiers in Ecology and Evolution* 8: 479.
- Reichman, O. J., M. B. Jones, and M. P. Schildhauer. 2011. “Challenges and Opportunities of Open Data in Ecology.” *Science* 331: 703–5.
- Rudenko, N., M. Golovchenko, L. Grubhoffer, and J. H. Oliver, Jr. 2011. “Updates on *Borrelia burgdorferi* sensu lato Complex with Respect to Public Health.” *Ticks and Tick-Borne Diseases* 2: 123–8.
- Sauer, J. R., K. L. Pardieck, D. J. Ziolkowski, Jr., A. C. Smith, M.-A. R. Hudson, V. Rodriguez, H. Berlanga, D. K. Niven, and W. A. Link. 2017. “The First 50 Years of the North American Breeding Bird Survey.” *The Condor: Ornithological Applications* 119: 576–93.
- Thibault, K. 2018. *TOS Protocol and Procedure: Breeding Landbird Abundance and Diversity*. NEON.DOC.014041vJ. Boulder, CO: NEON.
- Thibault, K., K. Tsao, Y. Springer, and L. Knapp. 2019. *TOS Protocol and Procedure: Small Mammal Sampling*. NEON.DOC.000481vL. Boulder, CO: NEON.
- Thorpe, A. S., D. T. Barnett, S. C. Elmendorf, E.-L. S. Hinckley, D. Hoekman, K. D. Jones, K. E. LeVan, C. L. Meier, L. F. Stanish, and K. M. Thibault. 2016. “Introduction to the Sampling Designs of the National Ecological Observatory Network Terrestrial Observation System.” *Ecosphere* 7: e01627.
- USDA, NRCS. 2022. The PLANTS Database. Greensboro, NC: National Plant Data Team. <https://plants.usda.gov>.
- Van Jaarsveld, A. S., J. C. Pauw, S. Mundree, S. Mecenero, B. W. Coetzee, and G. F. Alard. 2007. “South African Environmental Observation Network: Vision, Design and Status: SAEON Reviews.” *South African Journal of Science* 103: 289–94.
- Vellend, M., L. Baeten, I. H. Myers-Smith, S. C. Elmendorf, R. Beauséjour, C. D. Brown, P. De Frenne, K. Verheyen, and S. Wipf. 2013. “Global Meta-Analysis Reveals No Net Change in Local-Scale Plant Biodiversity over Time.” *Proceedings of the National Academy of Sciences* 110: 19456–9.
- Welti, E., A. Joern, A. M. Ellison, D. Lightfoot, S. Record, N. Rodenhouse, E. Stanley, and M. Kaspari. 2021. “Studies of Insect Temporal Trends Must Account for the Complex Sampling Histories Inherent to Many Long-Term Monitoring Efforts.” *Nature Ecology and Evolution* 5(5): 589–91. <https://doi.org/10.1038/s41559-021-01424-0>.
- Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific data* 3: 1–9.
- Worm, B., and D. P. Tittensor. 2018. *A Theory of Global Biodiversity (MPB-60)*. Princeton, NJ: Princeton University Press.

**How to cite this article:** Li, Daijiang, Sydne Record, Eric R. Sokol, Matthew E. Bitters, Melissa Y. Chen, Y. Anny Chung, Matthew R. Helmus, et al. 2022. “Standardized NEON Organismal Data for Biodiversity Research.” *Ecosphere* 13(7): e4141. <https://doi.org/10.1002/ecs2.4141>