Standardized NEON organismal data for biodiversity research

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

01 September, 2021

[1] Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, United States

[2] Center for Computation & Technology, Louisiana State University, Baton Rouge, LA, United States

[3] Department of Biology, Bryn Mawr College, Bryn Mawr, PA, United States

[4] Battelle, National Ecological Observatory Network (NEON), Boulder, CO, United States

[5] Institute of Arctic and Alpine Research (INSTAAR), University of Colorado Boulder, Boulder, CO, United States

[6] Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO, United States

[7] Departments of Plant Biology and Plant Pathology, University of Georgia, Athens, GA, United States

[8] Integrative Ecology Lab, Center for Biodiversity, Department of Biology, Temple University, Philadelphia, PA, United States

[9] St. Edward's University, Austin, Texas

[10] Department of Environmental Science and Management, Portland State University, Portland, OR, United States

[11] Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH, United States

[12] Translational Data Analytics Institute, The Ohio State University, Columbus, OH, United States

[13] Ecological Processes Branch, U.S. Army ERDC CERL, Champaign, IL, United States

[14] Department of Biological Sciences, DePaul University, Chicago, IL, United States

[15] Department of Environmental Science, Policy, and Management, University of California Berkeley, Berkeley, CA, United States

[16] School of Informatics, Computing and Cyber Systems, Northern Arizona University, Flagstaff, AZ, United States

[17] Environmental Data Initiative, University of Wisconsin-Madison, Madison, WI

[18] Department of Biological Sciences, Bridgewater State University, Bridgewater, MA, United States

[19] Department of Integrative Biology, Oregon State University, Corvallis, OR, United States

[20] Department of Integrative Biology, Michigan State University, East Lansing, MI, United States

[21] Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, MI, United States

† Equal contributions

‡ Corresponding authors: dli30@lsu.edu; srecord@brynmawr.edu; esokol@battelleecology.org

Open Research 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 a data package, which is available at Github (https://github.com/daijiang/neonDivData). Data were also permanently archived at the EDI data repository (https://portal-s.edirepository.org/nis/mapbrowse?scope=edi&identifier=190&revision=2). The code in the Supporting Information (CodeS1) is novel and will be available at Github upon acceptance.

Abstract: Understanding patterns and drivers of species distributions and abundances, 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 spans large spatial scales and long time periods. The National Ecological Observatory Network (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 provides 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 ecocomDP (ecological community data design pattern) format, and available through the ecocomDP 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.

Key words: NEON, Biodiversity, Organismal Data, Data Product, R, Data package, EDI

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 (Midgley and Thuiller 2005, Blowes et al. 2019). Such understanding is only possible through studies that address questions like: 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 and 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 1950's (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 (MacArthur and Wilson 1967, Hubbell 2001, Brown et al. 2004, Harte 2011) 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 (Vellend et al. 2013, Blowes et al. 2019, Li et al. 2020), but this work is challenged by the difficulty of integrating data from different studies and with varying limitations. Such limitations include: differing collection methods (methodological uncertainties); varying levels of statistical robustness; inconsistent handling of missing data; spatial bias; publication bias; and design flaws (Martin et al. 2012, Nakagawa and Santos 2012, Koricheva and Gurevitch 2014, Welti 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 and Hedges 1999).

Barriers to meta-analyses have been reduced in recent years to bring biodiversity research into the big data era (Hampton et al. 2013, Farley et al. 2018) 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 (Martin et al. 2012, Beck et al. 2014, Geldmann et al. 2016). 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, G Pricope et al. 2019). The standardized observational sampling of woody trees by the United States Forest Service's Forest Inventory and Analysis and of birds by the United States Geological Survey's Breeding Bird Survey have been ongoing across the United States since 2001 and 1966, respectively (Bechtold and Patterson 2005, Sauer et al. 2017), but cover few taxonomic groups. The Long Term Ecological Research Network (LTER) and Critical Zone Observatory (CZO) both are hypotheses-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 National Ecological Observatory Network (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 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 calculated for others; and data are in a raw form that often includes metadata unnecessary for biodiversity analyses. 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 and to advance macrosystem 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 ecocomDP 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 ecocomDP format through the R package ecocomDP (Smith et al. 2021) and more datasets are in the queue for processing into the ecocomDP format by EDI (O'Brien et al. 2021).

We standardized NEON organismal data into the ecocomDP format and all R code to process NEON data products can be obtained through the R package ecocomDP. 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. To facilitate the usage of the standardized datasets, we also developed an R data package, neonDivData. We refer to the input data streams provided by NEON as data products, whereas 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.

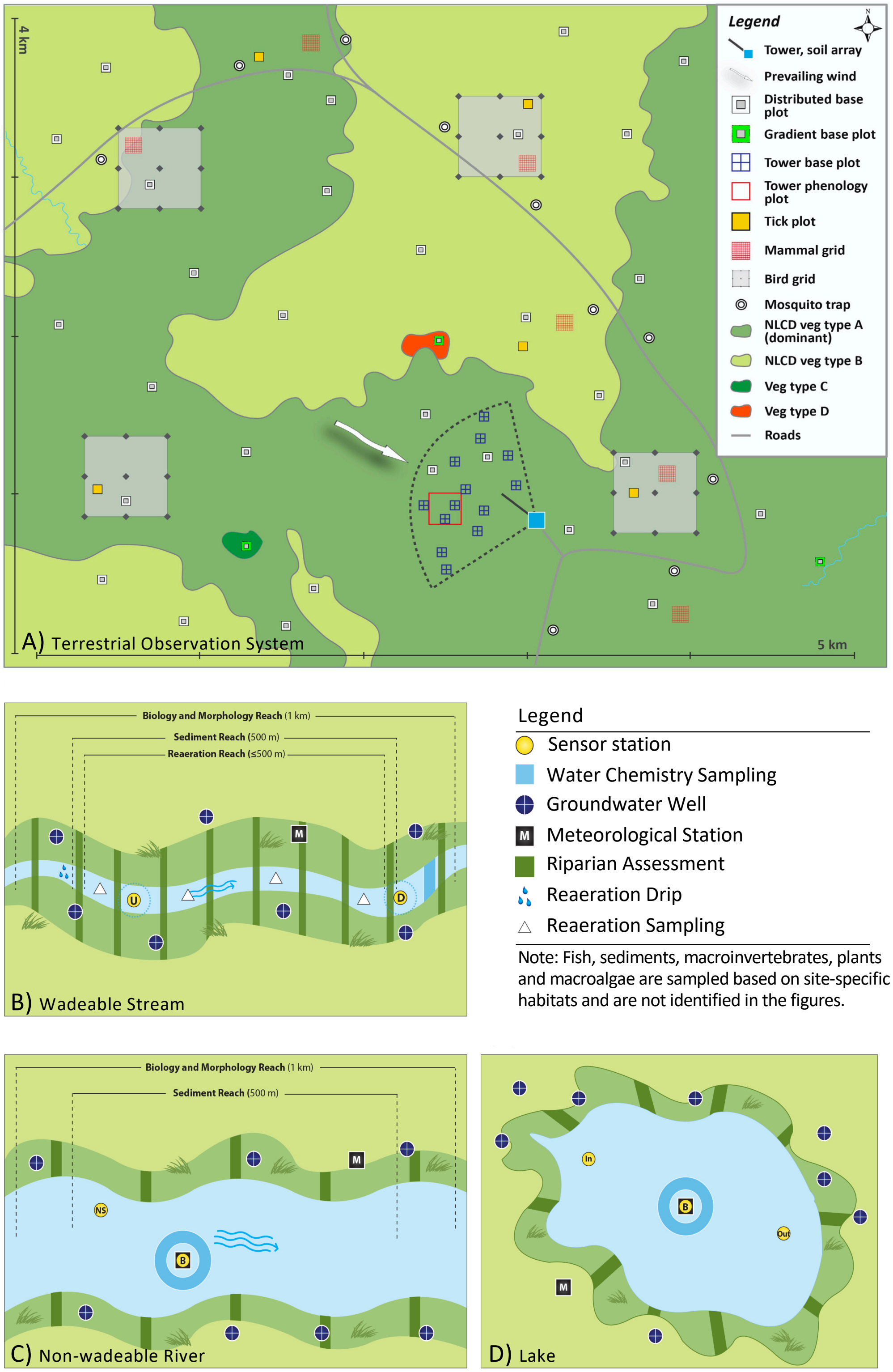
Materials and Methods (or how to standardize NEON organismal data)

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 regards to the NEON sampling design and decisions that were made as the data products presented in this paper were converted into the ecocomDP 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, etc.) 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 twelve taxonomic groups represented in eleven NEON data products produce standardized counts based on sampling effort, such as 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 by using the R ecocomDP package (Smith et al. 2021). To view a catalog of available NEON data products in the ecocomDP format, use ecocomDP::search_data("NEON"). To import data from a given NEON data product into your R environment, use ecocomDP::read_data(), and set the id argument to the selected NEON to ecocomDP mapping workflow (the "L0 to L1 ecocomDP workflow ID" in Table 1). This will return a list of ecocomDP 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 ecocomDP::flatten_ecocomDP() function. Second, 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 Issues log for each data product on the NEON Data Portal as well as the Readme text file that is included with NEON data downloads.

Terrestrial Organisms

Breeding Land Birds

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



Note: Fish, sediments, macroinvertebrates, plants and macroalgae are sampled based on site-specific habitats and are not identified in the figures.

Figure 1: Generalized sampling schematics for Terrestrial Observation System (A) and Aquatic Observation System (B-D) plots. For Terrestrial Observation System (TOS) plots, Distributed, Tower, and Gradient plots, and locations of various sampling regimes, are presented via symbols. For Aquatic Observation System (AOS) plots, Wadeable streams, Non-wadeable 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).

Table 1: Mapping NEON data products to ecocomDP formatted data packages with abundances *standardized* to observation effort. IDs in the L0 to L1 ecocomDP workflow ID columns were used in the R package ecocomDP to standardize organismal data. Notes: *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 abundances are beyond the scope of this paper; ** plants percent cover value NA represents presence/absence data only; *** incidence rate per number of tests conducted is reported for tick pathogens.

Taxon group	L0 dataset (NEON data product ID)	Version of NEON data used in this study	L0 to L1 ecocomDP workflow ID	Primary variable reported in ecocomDP observation table	Units
Algae	DP1.20166.001	https://doi.org/10.48443/3cvp-hw55 and provisional data	neon.ecocomdp.20166.001.001	cell density OR cells OR valves	cells/cm2 OR cells/mL
Beetles	DP1.10022.001	https://doi.org/10.48443/tx5f-dy17 and provisional data	neon.ecocomdp.10022.001.001	abundance	count per trap day
Birds*	DP1.10003.001	https://doi.org/10.48443/s730-dy13 and provisional data	neon.ecocomdp.10003.001.001	cluster size	count of individuals
Fish	DP1.20107.001	https://doi.org/10.48443/17cz-g567 and provisional data	neon.ecocomdp.20107.001.001	abundance	catch per unit effort
Herptiles	DP1.10022.001	https://doi.org/10.48443/tx5f-dy17 and provisional data	neon.ecocomdp.10022.001.002	abundance	count per trap day
Macroinvertebrates	DP1.20120.001	https://doi.org/10.48443/855x-0n27 and provisional data	neon.ecocomdp.20120.001.001	density	count per square meter
Mosquitoes	DP1.10043.001	https://doi.org/10.48443/9smm-v091 and provisional data	neon.ecocomdp.10043.001.001	abundance	count per trap hour
Plants**	DP1.10058.001	https://doi.org/10.48443/abge-r811 and provisional data	neon.ecocomdp.10058.001.001	percent cover	percent of plot area covered by taxon
Small mammals	DP1.10072.001	https://doi.org/10.48443/j1g9-2j27 and provisional data	neon.ecocomdp.10072.001.001	count	unique individuals per 100 trap nights per plot per month
Tick pathogens***	DP1.10092.001	https://doi.org/10.48443/5fab-xv19 and provisional data	neon.ecocomdp.10092.001.001	positivity rate	positive tests per pathogen per sampling event
Ticks	DP1.10093.001	https://doi.org/10.48443/dx40-wr20 and provisional data	neon.ecocomdp.10093.001.001	abundance	count per square meter
Zooplankton	DP1.20219.001	https://doi.org/10.48443/qzr1-jr79 and provisional data	neon.ecocomdp.20219.001.001	density	count per liter

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 five to ten 3 × 3 grids (Fig. 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-minute period, following a 2-minute 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. 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 (IMBCR) 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 ecocomDP 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, etc.). The field taxon_id in the R object data_bird with the neonDivData data package consists of the standard AOU 4-letter species code, although taxon_rank refers to eight 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 Beetles 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 et al. 2019b).

Beetle pitfall trapping begins when the temperature has been >4℃ 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 re-identified by an expert taxonomist. More details can be found in Hoekman et al. (2017) and LeVan et al. (2019b).

Pitfall traps and sampling methods are designed by NEON to reduce vertebrate bycatch (LeVan et al. 2019b). 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 morbund vertebrates are euthanized and collected along with deceased vertebrates. When ≥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 multi-step 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 being labelled "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 data set 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 non-expert taxonomists are more likely to misidentify non-native carabid species (Egli et al. 2020). Beetle abundances are 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 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 collectDate's. Second, there were a few instances in which plots within a site were set and collected on the same day, but have different EventID's. 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 beetles and herps bycatch were standardized to be the number of individuals captured per trap day.

Mosquitos

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 et al. 2019a). 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 et al. 2019a). Technicians deploy traps at all 10 dedicated mosquito trap locations per site. Traps remain out for a 24-hour period, or sampling bout, and bouts occur every two or four 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 eight hours after the trap was set, and one day-active sample, collected at dusk or ~16 hours after the trap was set. Thus, a 24-hour sampling bout yields 20 samples from 10 traps.

NEON collects mosquito specimens using Center for Disease Control (CDC) CO_2 light traps (LeVan et al. 2019a). 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 data sets. A CDC CO_2 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_2. Mosquitoes attracted to the CO_2 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 lab 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 as well as long-term archiving. Particularly rare or difficult to identify mosquito specimens are prioritized for DNA barcoding. More details can be found in LeVan et al. (2019a).

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_expertTaxonomistIDProcessed). 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 $\text{(individualCount/subsampleWeight) * totalWeight / trapHours}$. 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 is (1) nonvolant; (2) nocturnally active; (3) forages predominantly aboveground; and (4) has a mass >5 grams, but <~ 500-600 grams (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 m × 90 m trapping grids are set up in each terrestrial field site within the dominant vegetation type. Each 90 m × 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 m × 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 m^2 (20 m × 20 m) plots (Barnett 2019). Each multiscale plot is subdivided into four 100 m^2 (10 m × 10 m) subplots that each encompass one or two sets of 10 m^2 (3.16 m × 3.16 m) subplots within which a 1 m^2 (1 m × 1 m) subplot is nested. The percent cover of each plant species is estimated visually in the 1 m^2 subplots, while only species presences are documented in the 10 m^2 and 100 m^2 subplots.

To estimate plant percent cover by species, technicians record this value for all species in a 1 m^2 subplot (Barnett 2019). Next, the remaining 9 m^2 area of the associated 10 m^2 subplot is searched for the presence of species. The process is repeated if there is a second 1 and 10 m^2 nested pair in the specific 100 m^2 subplot. Next, the remaining 80 m^2 area is searched for the presence of species; data can be aggregated for a complete list of species present at the 100 m^2 subplot scale. Data for all four 100 m^2 subplots represent indices of species at the 400 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 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. 2014. 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 as well as 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 (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 m × 1 m scale also includes observations of abiotic and non-target species ground cover (i.e., soil, water, downed wood), so we removed records with divDataType as "otherVariables." We also removed records whose targetTaxaPresent is N (i.e., a non-target species). Additionally, for all spatial resolutions (i.e., 1 m^2, 10 m^2, and 100 m^2 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^2 scale, we incorporated all data from both the 1 m^2 and 10 m^2 scales for that subplot. Similarly, to obtain all records within a plot at the 400 m^2 scale, we included all data from that plot. Species abundance information was only recorded as area coverage within 1 m 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 et al. 2019c). Ticks are sampled by walking the perimeter of a 40 m × 40 m plot using a 1 m × 1 m drag cloth. Ideally, 160 meters are sampled (shortest straight line distance between corners), but the cloth can be dragged around obstacles if a straight line is not possible. 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 coincides with phenological milestones at each site, beginning within two weeks of the onset of green-up and ending within two weeks of vegetation senescence (LeVan et al. 2019c). Sampling bouts are only initiated if the high temperature on the two consecutive days prior to planned sampling was >0℃. Early season sampling is conducted on a low intensity schedule, with one sampling bout every six 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 three 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 et al. 2019c). 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 *Anaplasma 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) 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 lab 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 re-formatted 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 "lab 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 lab counts were available, they did not always agree (8% of records). In cases of disagreement, we generally used lab 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 lab, which resulted in higher field counts than lab counts. In this case, we assigned the additional un-identified individuals (e.g., the difference between the field and lab 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 lab 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 lab or field, therefore this process could be ignored by users whose analyses are not sensitive to exact counts.

The tick pathogen NEON data product (DP1.10092.001) 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 *Borrelia 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 number 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. 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 lab for sorting and identification to lowest possible taxon (e.g., genus or species) and counts of each taxon per size are made to the nearest mm.

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 mm 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 lab) 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 as well as 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 non-stratified lakes and non-wadeable 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 microalgae 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, 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.

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 =

'cellsperBottle').

We corrected sample units of cellsperBottle to density (Parker and Vance 2020). First, we summed the preservative volume and the lab'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 lab 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, divided by benthic area sampled), in square meters. The final abundance units are either cells/mL (phytoplankton and seston samples) or cells/m^2 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 sampleID's per parentSampleID with one for ash-free dry mass (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 eco-climatic domains, occuring in a total of 23 lotic (stream) and five lentic (lake) sites. In lotic sites, up to 10 non-overlapping reaches, each 70 to 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 five days and with a minimum two-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 m × 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 minutes between successive passes. Electrofishing begins within 30 minutes of sunset and ceases within 30 minutes 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 hours 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 fsh_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 into 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 that 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 and 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 lab for sorting and identification to lowest possible taxon (e.g., genus or species) and counts of each taxon per size are made to the nearest mm.

Data Wrangling Decisions The NEON zooplankton data product (DP1.20219.001) consists of dataframes for taxonomic identification and related field data (Parker and Scott 2020). Zooplankton in NEON samples are identified at contracting labs 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 (temporary link, which will be finalized upon acceptance: https://portal-s.edirepository.org/nis/mapbrowse?scope=edi&identifier=190&revision=2). 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 ecocomDP and modify them for their own purposes.

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 the 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 were 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 were saved in the neon_location

Table 2: Summary of data products included in this study (as of 01 September, 2021). 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	N species	N sites	Start date	End date
Algae	data_algae	1946	33	2014-07-02	2019-07-15
Beetles	data_herp_bycatch	756	47	2013-07-03	2020-10-13
Birds	data_bird	541	47	2015-05-13	2020-07-20
Fish	data_fish	147	28	2016-03-29	2020-12-03
Herptiles	data_herp_bycatch	128	41	2014-04-02	2020-09-29
Macroinvertebrates	data_macroinvertebrate	1330	34	2014-07-01	2020-08-12
Mosquitoes	data_mosquito	128	47	2014-04-09	2020-06-16
Plants	data_plant	6197	47	2013-06-24	2020-10-23
Small mammals	data_small_mammal	145	46	2013-06-19	2020-11-20
Tick pathogens	data_tick_pathogen	12	15	2014-04-17	2018-10-03
Ticks	data_tick	19	46	2014-04-02	2020-10-06
Zooplankton	data_zooplankton	157	7	2014-07-02	2020-07-22

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 were 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 (Fig. 2). To show how easy it is to get site level species richness, we presented the code used to generate the data for Fig. 2 as CodeS1 in the supporting information.

Figure 2 shows the utility of the data package for exploring macroecological patterns at the NEON site level. One of the most well known and studied macroecological patterns is the latitudinal biodiversity gradient, wherein sites are more species 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 data packages presented in this paper enable investigation of effects of taxonomic

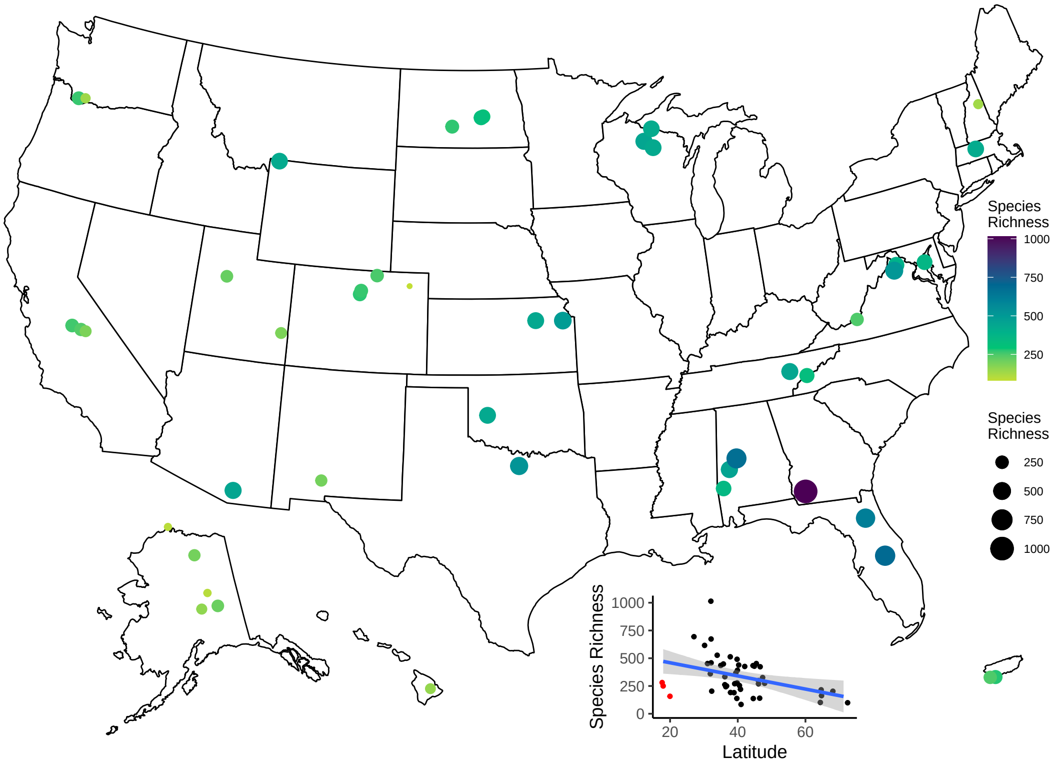


Figure 2: Plant species richness mapped across NEON terrestrial sites. The inset scatterplot shows latitude on the x-axis and species richness on the y-axis, with red points representing sites in Puerto Rico and Hawaii.

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 and 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 diversity indices (Jost 2006) for beetles at the Oak Ridge National Laboratory (ORNL) NEON site for data subsetted at the genus, species, and subspecies level. To quantify biodiversity, we used Jost 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 Fig. 3). For the ORNL beetle data, Jost 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, 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 genus.

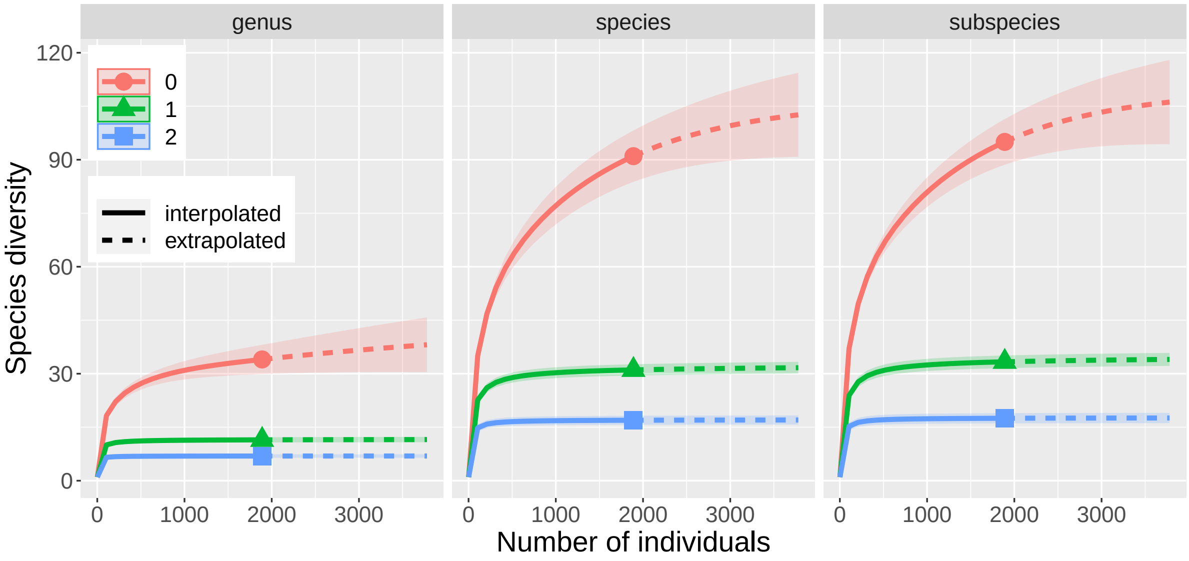


Figure 3: Rarefaction of beetle abundance data from collections made at the Oak Ridge National Laboratory (ORNL) National Ecological Observatory Network (NEON) site from 2014-2020 generated using the beetle data package presented in this paper and the iNEXT package in R (Hsieh et al. 2016) based on different levels of taxonomic resolution (i.e., genus, species, subspecies). Different colors indicate Jost Indices with differing values of q (Jost 2006).

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 and 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 data sets 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 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.

There are some important notes about the data package we provided. First, 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 IDs level. Such information can be useful if we are interested in questions such as species-to-genus ratio or species rarefaction curves at different taxonomic levels (e.g., Fig. 3). 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.' vs '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 into estimates of biodiversity 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. Second, we standardized species abundance measurements to make them comparable across different sampling events within each taxonomic group (Table 1). Such standardization is critical to study and compare biodiversity. And finally, 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 abundances (DP1.20277.001), metagenome sequences (DP1.20279.001), marker gene sequences (DP1.20280.001), and community composition (DP1.20086.001); surface water microbe abundances (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. in prep.

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 data sets 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.

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 synthesis of NEON data with data from other efforts (e.g., LTER, CZO, Ameriflux, the International LTER, National Phenology Network, 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 data sets will help in this endeavor because they naturally promote the collection of provenance as data are collated into derived products (Reichman et al. 2011, O'Brien et al. 2021). 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.

Acknowledgement

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 award under Grant No. 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 S.R., NSF DEB 1926567 to P.L.Z., NSF DEB 1926598 to M.A.J, and NSF DEB 1926341 to J.M.L.. Comments from NEON staff (Katie LeVan, Dylan Mpnahan, Sata Paull, Dave Barnett, Sam Simkin), Margaret O'Brien and Tad Dallas greatly improved this work. The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle Memorial Institute. This material is based in part upon work supported by the National Science Foundation through the NEON Program.

Reference

Baker, J. R., D. V. Peck, and D. W. Sutton. 1997. Environmental monitoring and assessment program surface waters: Field operations manual for lakes. US Environmental Protection Agency, Washington.

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.

Barnett, D. 2019. TOS protocol and procedure: DIV - plant diversity sampling. NEON.DOC.014042vK. NEON (National Ecological Observatory Network).

Barnett, D. T., P. B. Adler, B. R. Chemel, P. A. Duffy, B. J. Enquist, J. B. Grace, S. Harrison, R. K. Peet, D. S. Schimel, T. J. Stohlgren, and others. 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. 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 19:10–15.

Blowes, S. A., S. R. Supp, L. H. Antão, A. Bates, H. Bruelheide, J. M. Chase, F. Moyes, A. Magurran, B. McGill, I. H. Myers-Smith, and others. 2019. The geography of biodiversity change in marine and terrestrial assemblages. Science 366:339–345.

Brown, J. H., J. F. Gillooly, A. P. Allen, V. M. Savage, and G. B. West. 2004. Toward a metabolic theory of ecology. Ecology 85:1771–1789.

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. NEON (National Ecological Observatory Network).

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 (dp1.20120.001). Revision b. NEON (National Ecological Observatory Network).

Curtis, J. T. 1959. The vegetation of wisconsin: An ordination of plant communities. University of Wisconsin Pres.

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 68:563–576.

Farrell, K. J., and C. C. Carey. 2018. Power, pitfalls, and potential for integrating computational literacy into undergraduate ecology courses. Ecology and evolution 8:7744–7751.

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 22:1139–1149.

G Pricope, N., 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.

Gurevitch, J., and L. V. Hedges. 1999. Statistical issues in ecological meta-analyses. Ecology 80:1142–1149.

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 11:156–162.

Harte, J. 2011. Maximum entropy and ecology: A theory of abundance, distribution, and energetics. OUP Oxford.

Heffernan, J. B., P. A. Soranno, M. J. Angilletta Jr, L. B. Buckley, D. S. Gruner, T. H. Keitt, J. R. Kellner, J. S. Kominoski, A. V. Rocha, J. Xiao, and others. 2014. Macrosystems 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, C. B. Knisley, J. R. LaBonte, J. Lundgren, and others. 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 (h ill numbers). Methods in Ecology and Evolution 7:1451–1456.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography (mpb-32). Princeton University Press.

Hutchinson, G. E. 1959. Homage to santa rosalia or why are there so many kinds of animals? The American Naturalist 93:145–159.

Jensen, B., S. Parker, and J. R. Fischer. 2019a. AOS protocol and procedure: Fish sampling in wadeable streams. NEON.DOC.001295vF. NEON (National Ecological Observatory Network).

Jensen, B., S. Parker, and J. R. Fischer. 2019b. AOS protocol and procedure: Fish sampling in lakes. NEON.DOC.001296vF. NEON (National Ecological Observatory Network).

Jones, J., P. Groffman, J. Blair, F. Davis, H. Dugan, E. Euskirchen, S. Frey, T. Harms, E. Hinckley, M. Kosmala, and others. 2021. Synergies among environmental science research and monitoringnetworks: A research agenda. Earth's Future:e2020EF001631.

Jost, L. 2006. Entropy and diversity. Oikos 113:363–375.

Keller, M., D. S. Schimel, W. W. Hargrove, and F. M. Hoffman. 2008. A continental strategy for the national ecological observatory network. The Ecological Society of America: 282-284.

Koricheva, J., and J. Gurevitch. 2014. Uses and misuses of meta-analysis in plant ecology. Journal of Ecology 102:828–844.

LeVan, K. 2020a. NEON user guide to ground beetles sampled from pitfall traps (dp1.10022.001).version c. NEON (National Ecological Observatory Network).

LeVan, K. 2020b. NEON user guide to mosquitoes sampled from co2 traps (dp1.10043.001) and mosquito-borne pathogen satatus (dp1.10041.001).version c. NEON (National Ecological Observatory Network).

LeVan, K., S. Paull, K. Tsao, D. Hoekman, and Y. Springer. 2019a. TOS protocol and procedure: MOS - mosquito sampling. NEON.DOC.014049vL. NEON (National Ecological Observatory Network).

LeVan, K., N. Robinson, D. Hoekman, and K. Blevins. 2019b. TOS protocol and procedure: Ground beetle sampling. NEON.DOC.014041vJ. NEON (National Ecological Observatory Network).

LeVan, K., K. Thibault, K. Tsao, and Y. Springer. 2019c. TOS protocol and procedure: Tick and tick-borne pathogen sampling. NEON.DOC.014045vK. NEON (National Ecological Observatory Network).

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.

Linnaeus, C. 1758. Systema naturae. Stockholm Laurentii Salvii.

MacArthur, R. H., and E. O. Wilson. 1967. The theory of island biogeography. 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 167:638–641.

Monahan, D., B. Jensen, S. Parker, and C. Scott. 2020. NEON user guide to fish electrofishing, gill netting, and fyke netting counts (dp1.20107.001). Revision b. NEON (National Ecological Observatory Network).

Moulton II, S. R., 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. Geological Survey (US).

Nakagawa, S., and E. S. Santos. 2012. Methodological issues and advances in biological meta-analysis. Evolutionary Ecology 26:1253–1274.

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.

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 3:21–29.

Parker, S. 2019. AOS protocol and procedure: INV - aquatic macroinvertebrate sampling. NEON.DOC.003046vE. NEON (National Ecological Observatory Network).

Parker, S. 2020. AOS protocol and procedure: ALG - periphyton and phytoplankton sampling. NEON.DOC.003045vE. NEON (National Ecological Observatory Network).

Parker, S., and C. Roehm. 2019. AOS protocol and procedure: ZOO - zooplankton sampling in lakes. NEON.DOC.001194. NEON (National Ecological Observatory Network).

Parker, S., and C. Scott. 2020. NEON user guide to aquatic zooplankton collection (dp1.20219.001). Revision b. NEON (National Ecological Observatory Network).

Parker, S., and T. Vance. 2020. NEON user guide to periphyton and phytoplankton collection (dp1.20166.001). Revision c. NEON (National Ecological Observatory Network).

Pavlacky Jr, D. C., 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., Herlihy, A. T., Hill, B. H., Hughes, R. M., Kaufmann, P. R., Klemm, D. J., Lazorchak, J. M., McCormick, F. H., Peterson, S. A., Ringold, P. L., Magee, T., and M. R. and Cappaert. 2006. Environmental monitoring and assessment program — surface waters: Western pilot study field operations manual for wadeable streams. US Environmental Protection Agency, Washington.

Pritt, B. S., M. E. Allerdice, L. M. Sloan, C. D. Paddock, U. G. Munderloh, Y. Rikihisa, T. Tajima, S. M. Paskewitz, D. F. Neitzel, D. K. H. Johnson, and others. 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.

Rainio, J., and J. Niemelä. 2003. Ground beetles (coleoptera: Carabidae) as bioindicators. Biodiversity & Conservation 12:487–506.

Ralph, C. J. 1993. Handbook of field methods for monitoring landbirds. 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 41:1718–1727.

Record, S., N. M. Voelker, P. L. Zarnetske, N. I. Wisnoski, J. D. Tonkin, C. Swan, L. Marazzi, N. Lany, T. Lamy, A. Compagnoni, and others. 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 331:703–705.

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 2:123–128.

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 119:576–593.

Smith, C., E. Sokol, and M. O'Brien. 2021. EcocomDP: Work with datasets in the ecological community design pattern.

Thibault, K. 2018. TOS protocol and procedure: Breeding landbird abundance and diversity. NEON.DOC.014041vJ. NEON (National Ecological Observatory Network).

Thibault, K., K. Tsao, Y. Springer, and L. Knapp. 2019. TOS protocol and procedure: Small mammal sampling. NEON.DOC.000481vL. NEON (National Ecological Observatory Network).

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 n ational e cological o bservatory n etwork t errestrial o bservation s ystem. Ecosphere 7:e01627.

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 110:19456–19459.

Welti, E., A. Joern, A. M. Ellison, D. Lightfoot, S. Record, N. Rodenhouse, E. Stanley, and M. Kaspari. 2021. Meta-analyses of insect temporal trends must account for the complex sampling histories inherent to many long-term monitoring efforts. Nature Ecology and Evolution.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, and others. 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 University Press.