# Tidy NEON organismal data for biodiversity research

Daijiang Li[1,2], Sydne Record[3], Eric Sokol[4], Matthew E. Bitters, Melissa Y. Chen, Anny Y. Chung, Matthew Helmus, Ruvi Jaimes, Lara Jansen, Marta A. Jarzyna, Michael G. Just, Jalene M. LaMontagne, Brett Melbourne, Wynne Moss, Kari Norman, Stephanie Parker, Natalie Robinson, Bijan Seyednasrollah, Colin Smith, Sarah Spaulding, Thilina Surasinghe, Sarah Thomsen, Phoebe Zarnetske

18 January, 2021

## Contents



**Abstract**: Understanding patterns and drivers of the distribution and abundance of species, and thus biodiversity, is a core goal of ecology. Despite the advances made over the past decades on pursuing this goal, we are currently limited by the lack of standardized high quality empirical data across large spatial scales and long time periods. The National Ecological Observatory Network (NEON) provides such a database by conducting robust organismal sampling of several sentinel taxonomic groups across 81 sites distributed across the United States. NEON is now fully operational and will continue for 30 years, offering a unique opportunity to advance biodiversity research. To maximize the potential of this opportunity, however, it is critical to

lower the hurdles for the usage of NEON organismal data. Here, we took the first step to tidy the NEON organismal data to facilitate its usage for biodiversity research. We briefly summarized sampling designs for major taxonomic groups and documented our decisions to wrangle their corresponding NEON organismal data. All codes to perform such data wrangling are freely available online, allowing contributions by others and regular updates when new data are available. Finally, we demonstrated the usage of the tidied data with two simple examples. By providing a tidied and harmonized data product, it is our hope to advance biodiversity research based on NEON organismal data in a comparable and reproducible way.

**Key words**: NEON, Biodiversity, Organismal Data, Data Product

# Introduction (or why tidy NEON organismal data)

A central goal of ecology is to understand the patterns and processes of biodiversity, which is particularly important in an era of rapid global environmental change (Midgley and Thuiller 2005, Blowes et al. 2019). Such understanding comes from addressing 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 geographies? What are the local and landscape scale drivers of community structure? How and why do biodiversity patterns change over time? Answers to such questions are essential to understanding, managing, and conserving biodiversity and the ecosystem services it influences.

Biodiversity research has a long history (Worm and Tittensor 2018), beginning with major scientific expeditions (e.g. Alexander von Humboldt, Charles Darwin) that were undertaken to explore global biodiversity after the establishment of Linnaeus's Systema Naturae (Linnaeus 1758). Modern biodiversity research dates back to the 1950s (Curtis 1959, Hutchinson 1959) and aims to quantify patterns of species diversity and describe mechanisms underlying its heterogeneity. Since the beginning of this line of research, major theoretical breakthroughs (MacArthur and Wilson 1967, Hubbell 2001, Brown et al. 2004) have advanced our understanding

of potential mechanisms causing and maintaining biodiversity.

Modern empirical studies that test mechanisms, however, have been largely constrained to local or regional scales, and focused on one or a few specific taxonomic groups. Despite such constraints, field ecologists have compiled unprecedented numbers of observations, which support research into generalities through syntheses and meta-analyses (Vellend et al. 2013, Blowes et al. 2019, Li et al. 2020). Such work is challenged, however, by the difficulty of bringing together data from different studies and with varying limitations, including: 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). 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). This has been remedied in recent years by large efforts to digitize museum and herbarium specimens (e.g., iDigBio), successful community science programs (e.g., iNaturalist, eBird), and advances in technology (e.g., remote sensing, automated acoustic recorders) that together bring biodiversity research into the big data era (Hampton et al. 2013, Farley et al. 2018). Yet, each of these comes with its own limitations. For example, museum/herbarium specimens and community science records are incidental (thus, unstructured in terms of the sampling design) and show obvious geographic and taxonomic biases (Martin et al. 2012, Beck et al. 2014, Geldmann et al. 2016); remote sensing approaches can cover large spatial scales, but may be of low spatial resolution and unable to reliably penetrate vegetation canopy (Palumbo et al. 2017, G Pricope et al. 2019). Overall, our understanding of biodiversity and ability to test for mechanisms is currently limited by the lack of standardized high quality and open-access data across large spatial scales and long time periods.

Long running coordinated research networks address some of these issues of standardized sampling across space and time for particular taxa across. For instance, 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 across North America since 1966, respectively (Bechtold and Patterson 2005, Sauer et al. 2017). The Long Term Ecological Research Network (LTER) consists of 28 sites that provide long term (up to 40 years) observational and

experimental datasets for a diverse set of ecosystems (i.e., terrestrial, aquatic, and marine) and taxa (e.g., microbes, animals, and plants). However, there is no standardization in the design and data collections across LTER sites (Jones et al. 2021).

The recently established National Ecological Observatory Network (NEON) builds off of these efforts by providing continental-scale observations to be collected using standardized methods over 30 years on various taxonomic groups with the goal of providing open access datasets broadly aimed at enabling better understanding of how U.S. ecosystems change through time (Keller et al. 2008). Data collected by NEON include observations and field surveys, automated instrument measurements, airborne remote sensing surveys, and archival samples that characterize plants, animals, soils, nutrients, freshwater and atmospheric conditions. Data are collected at 81 field sites across both terrestrial and freshwater ecosystems across the United States and are slated to continue for 30 years. These data provide a unique opportunity for advancing biodiversity research because consistent data collection protocols and the long-term nature of the observatory ensure sustained data availability and directly comparable measurements across locations. Spatio-temporal patterns in biodiversity, and the causes of changes to these patterns, can thus be confidently assessed and analyzed using NEON data.

NEON data are designed to be maximally useful to ecologists by aligning with FAIR (findable, accessible, interoperable, and reusable) principles (Wilkinson et al. 2016), but there are still hurdles to overcome in the reproducible use of NEON organismal data for biodiversity research. For example, different NEON data products use different field names for similar measurements and some include sampling unit information while units must be calculated for others. NEON organismal data products provide lots of raw data, but most of these provided raw data are unnecessary to calculate biodiversity measurements. Therefore, users need to dive into the comprehensive documentation to better understand the organismal datasets, to extract the relevant essential variables, and to take additional steps to wrangle the data to quantify biodiversity (e.g., clean datasets, change the data formats to feed the data to statistical programs, etc.). Such processes can be very time consuming and the path to a standard data format is different and not always obvious for each NEON organismal data product. Thus, although NEON organism data are collected in a standardized fashion, the data wrangling decisions made leading up to an analysis could be different depending on the researcher, even if the research question

necessitates the calculation of similar community level biodiversity metrics. Ultimately, these subtle differences from study to study could make synthesis of NEON enabled science challenging despite the standardization of the data collection itself. A data product that simplifies and standardizes various NEON organismal datasets can remove such hurdles, enhance the interoperability and reusability, and facilitate wider usage of NEON organismal datasets for biodiversity research.

Ideally, such a data product could also interface well with organismal data sets from other coordinated research networks and studies to promote synthesis and advance macrosystem biology beyond NEON data (Record et al. 2020). A recent effort that brought together members of the LTER network, the Environmental Data Initiative (EDI), and NEON staff resulted in the conceptualization of a data design pattern for analysis ready community level organismal data known as ecocomDP (O'Brien et al. In prep). The basic premise of ecocomDP is to maintain different dataset "levels" as data are transformed for analyses. Level 0 (L0) data are incoming or raw. Level 1 (L1) data is the L0 data transformed to the ecocomDP model comprised at a minimum of tables representing observations, sampling locations, and taxonomic information. Level 2 (L2) data represent data that have been further transformed for specific research studies. Thus far, >70 LTER organismal datasets have been harmonized to the L1 ecocomDP format and more data sets are in the queue for processing into the ecocomDP format by EDI.

Given the potential of the ecocomDP format to transform biodiversity synthesis research efforts, our goal is to provide a standardized "tidy version" of NEON organismal datasets in the ecocomDP format for biodiversity research. Users can download the tidy data from the R package `ecocomDP` (CITATION), which will be maintained and updated when new data is available from the NEON portal. Our hope is to standardize formats across NEON data products and substantially reduce data cleaning times for the community of ecological researchers, and to facilitate the use of NEON data to advance understanding of Earth's biodiversity.

# Materials and Methods (or how to tidy NEON organismal data)

There are an overwhelming number of details to consider when starting to use NEON organismal data. 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 format. While the methodological sections below are specific to particular taxonomic groups, there are some general points that we came across while pouring over NEON data product documentation and through discussion with NEON staff apply to all NEON organismal data products. For instance, across NEON organismal data products, the term 'sp.' refers to a single morphospecies whereas the term 'spp.' refers to more than one morphospecies. This is a key point to consider for community ecology biodiversity analyses because it may add uncertainty into estimates of biodiversity metrics such as species richness. We also recognize that NEON data collection will continue well after this paper is published and new changes to data collection or the way that data are treated by the NEON data products groups may vary over time. Issues log (e.g. number of traps change for beetles )

## Terrestrial Organisms

### Breeding Land Birds

**NEON Sampling Design** Landbirds are surveyed with point counts during the breeding season in each of the 47 terrestrial sites, co-located with distributed plots whenever possible (Fig. 1). Breeding landbirds are "smaller birds (usually exclusive of raptors and upland game birds) not usually associated with aquatic habitats" (Ralph 1993). At NEON sites, one sampling bout occurs per breeding season at large sites, and two sampling bouts occur at smaller sites. Point counts occur either within randomly distributed individual points or within bird grids at each site in representative (dominant) vegetation. At large NEON sites, 5-15 grids are sampled with nine point count locations each, where grid centers are co-located with distributed base plot centers,

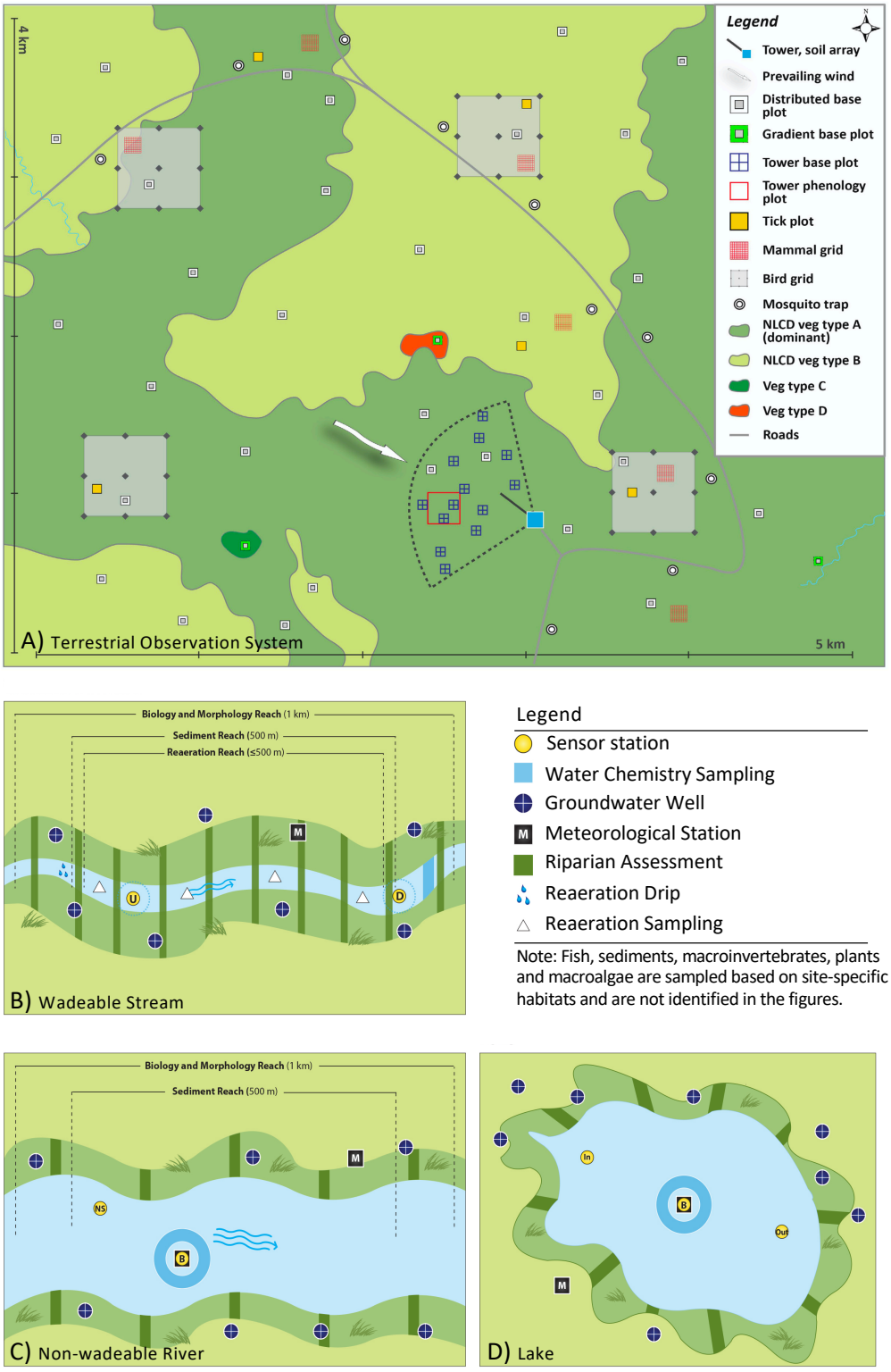


Figure 1: Generalized sampling schematics for Terrestrial Observation System (A) and Aquatic Observation System (B-D) plots. For Terrestrial Observation System (TOS) plots, Distributed, Tower, and Gradient plots, and locations of various sampling regimes, are presented via symbols. For Aquatic Observation System (AOS) plots, Wadeable stream, Non-wadeable stream, and Lake plots are shown in detail, with locations of sensors and different sampling regimes presented using symbols. Panel A was originally published by Thorpe et al. (2016).

if possible. If small sites only allow five grids, a stratified random sample maintains 250 m minimum separation between point count locations and point counts occur at the southwest corner of the 5-25 distributed base plots.

The breeding season month, which defines the timing of sampling, varies somewhat by site but always occurs in the spring. Most species observed are diurnal and include both resident and migrant species. Early in the morning observers conduct point counts wherein the observer tracks each minute. 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 during a 6-minute period after a 2-minute acclimation period. To enable subsequent modeling of detectability, additional data collected during the point counts include: weather, distances from observers to birds, and the detection methods. 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).

To protect species of concern, their taxonomic IDs are 'fuzzed.' This means the data are provided with a taxonomic identification at one higher taxonomic level than where the protection occurs. 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.

**Data Wrangling Decisions** Bird point count data ('DP1.10003.001'), consist 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 second data frame contains additional location information such as latitude longitude coordinates and environmental conditions during the time of the observations. It is relatively straightforward to prepare the bird point count data for biodiversity research. We first combined both data frames into one and then removed columns that are likely not needed (e.g., laboratory names, publication dates, etc.).

The sampling protocol has evolved over time, so users are advised to check whether the samplingProtocolVersion fits their data requirements and subset as necessary. Zero counts could be excluded if desired by subsetting on the field targetTaxaPresent before proceeding. The

field taxonID consists of the standard AOU 4-letter species code, although taxonRank 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 (taxonID = UNBI), where no further details are available below the class level (Aves sp.).

## Ground Beetles and Herp Bycatch

**NEON Sampling Design** Each site is sampled with pitfall traps, with 10 separate distributed plots at each site and four pitfall traps at each plot initially - placed in the ground along the cardinal directions of the distributed plot boundary. This equates to a total of 40 pitfall 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. Sampling 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. After collection, the samples are separated into carabid species and bycatch, with bycatch archived at either the trap (vertebrate) or plot (invertebrate) level. Carabid samples and vertebrate bycatch 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).

Pitfall traps and sampling methods are designed by NEON to reduce vertebrate bycatch Hoekman et al. (2017). 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 or more individuals of a vertebrate species are collected, cumulatively, within a single plot, NEON may initiatie 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, due to these active efforts to reduce vertebrate bycatch, it is unclear how well these pitfall traps capture herp community structure and diversity and users of the herp bycatch

data we wrangle and provide here should be aware of these limitations.

**Data Wrangling Decisions** Beetle samples are identified at multiple levels of expertise. Beetles are first identified by the sorting technician and then the pinning technician. Identifications of more difficult specimens are additionally verified by an expert taxonomist. Whenever available, expert identification is used for a sample. For example, if taxonomic delineation between NEON staff and multiple expert taxonomist identifications do not agree, then the consensus expert taxonomist delineation is recorded in the data portal. 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 on the sorted sample, by NEON technicians, and are not preserved across the different levels of identification. For example, a sample of 15 individuals identified during the sorting phase may be passed to a pinning technician, who then identifies five different species within that sample. The pinning technician does not back annotate the sorted sample to identify which individuals are which species, or go through and re-identify the rest of the individuals in the sample. Without this, we have assumed that all individuals in the sorted sample that were not positively identified by an expert were correctly identified in the original sample by NEON technicians. Hence, the abundance for a newly identified species is one, and the abundance for the originally identified species for the sample is the original abundance minus the individuals expertly identified as a different species.

Sometimes there are more individuals identified by pinning technicians or experts than were counted in the original sorted sample, so the count has been updated in the dataset. There are also 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.

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 account 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 are counted 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.

## Mosquitos

**NEON Sampling Design** Mosquito specimens are collected at 47 terrestrial sites across all NEON domains. Traps are distributed throughout the site according to a stratified-random spatial design used for all Terrestrial Observation System sampling, and are typically located within 30m of a road to facilitate expedient sampling. NEON collects mosquito specimens using the Center for Disease Control (CDC) $CO_2$ light traps. These traps have been used by other public health and mosquito-control agencies for a half-century, which allows NEON mosquito data to be used across field sites and in combination 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.

Mosquito monitoring is divided into field season and off-season sampling. Off-season sampling takes place weekly at core sites, and begins after three consecutive zero-catch field sampling bouts at the core site. The goal of off-season sampling is to rapidly determine when the next field season should begin and to provide mosquito phenology data throughout the lifetime of the observatory. During the off season, overnight sampling occurs weekly at three dedicated mosquito plots spread throughout the terrestrial core sites for each domain while temperatures are >10 ℃. Traps are deployed at dusk and checked the following dawn. Field season sampling begins when the first mosquito is detected during off season sampling.

Technicians collect samples every two weeks at core terrestrial sites and every four weeks at relocatable terrestrial sites. Sampling occurs at 10 dedicated mosquito plots at each site over a 24-hour period, or one sampling bout. During the sampling bout, traps are serviced twice and yield one night-active sample, taken at dawn or about eight hours after the trap is set, and one day-active sample, taken at dusk or ~16 hours after the trap is set. Thus, a 24-hour sampling bout yields 20 "samples" from 10 traps.

Following field collection, NEON's field ecologists process, pack up 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 Hoekman et al. (2016).

**Data Wrangling Decisions Data Wrangling Decisions** Mosquito data are mainly stored in 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 important information about location, collection date, and sample or subsample ID for all data frames. We

then merged all four data frames into one while checked carefully that they are merged correctly. In the merged data frame, we only kept records for target taxa (i.e., $\mathrm{targetTaxaPresent} == \text{"Y"}$) and no known compromised sampling condition (i.e., $\mathrm{sampleCondition} == \text{"No known compromise"}$). We further removed a small number of records with species identified only to the family level; all remaining records were identified at least at the genus level. We estimated the total individual count for each species within a trap as $\mathrm{individualCount} * (\mathrm{totalWeight} / \mathrm{subsampleWeight})$. We then removed columns that likely will not 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 < about 500-600 grams. 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. Multiple 90 m x 90 m trapping grids are set up in each terrestrial field site within the dominant vegetation type. Each 90 m x 90 m trapping grid contains 100 traps placed in a pattern with 10 rows and 10 columns set 10 m apart. Three 90 m x 90 m grids per site are designated pathogen grids and the remainder are designated diversity grids. Small mammal sampling occurs in bouts, with a bout comprised of three consecutive (or nearly consecutive) nights of trapping, and is based on the lunar calendar, with timing of sampling constrained to occur within 10 days before or after the new moon. The number of bouts per year is determined by site type, and most sites contain six bouts per year.

**Data Wrangling Decisions** In the data presented, 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), if they represented dead animals (fate = 'dead') or escaped animals (fate = 'escaped'), or bycatch (fate = 'nontarget,' i.e., non-target

species). Records for recaptured individuals were also removed. However, we kept empty traps as they contain information about sampling efforts, which can be useful for some studies.

## Soil Microbes

**NEON Sampling Design** Soil samples are collected at ten 40 x 40 m2 NEON plots per site. Four plots are within the tower airshed (tower plots), and six plots are distributed across the landscape (gradient plots). At each sampling time point, soils are sampled from three of the four subplots, and one sample collected from a randomly-generated XY coordinate location within each subplot. At each sampling location, soils are taken at the surface horizon most years, but from both organic and mineral horizons every five years during coordinated microbe/biogeochemistry bouts. Most sites, except for the boreal/arctic sites, are sampled three times a year, once at peak vegetation greenness and two other times bracketing that period. This results in ~10 plots * 3 locations * 1 or 2 horizons * 3 periods = 90 - 180 soil samples per site per year for most sites. Samples for microbial biomass, composition, and metagenomics are stored on dry ice and shipped to an external lab (variable depending on year) for downstream processing.

**Data Wrangling Decisions** Unlike other NEON biodiversity data, the soil microbial datasets require significant pre-processing to go from raw sequence data to a community matrix, and the exact bioinformatics methods will vary depending on use case. Briefly, major decisions during this process will depend on whether users are working with fungal (ITS) or bacterial (16S) data, if the goal is to maximize read quality and taxonomic resolution vs. number of reads retained through the quality filter process, and whether to remove or retain reverse complement reads for a merged sequence. The full description of a suggested bioinformatics pipeline, how to run sensitivity analyses on user-defined parameters, accompanying code, and vignettes are described in Qin *et al* in this issue. At the end of the suggested bioinformatics pipeline, users will have a phyloseq object, which is a commonly-used format for sequence-based analysis software. The phyloseq object will contain a table of ASV (amplicon sequence variant) sequences, a table of taxonomic assignments, and soil chemical and physical data associated with the same sample locations and sampling bouts.

## Terrestrial Plants

**NEON Sampling Design** NEON plant diversity plots are sampled during one or two bouts per year, and are 400 m x 400 m in size. Sampling is done using a nested design, where the entire plot is first subdivided into 4 100 m x 100 m subplots. For each subplot, one or more 1 m x 1 m nested subplots are then sampled; species coverages within the 1 $m^2$ area are estimated visually. Next presence/absence of plants is recorded in one or more 10 m x 10 m subplots, inside of which the finer resolution subplots are located. Finally, the 100 m x 100 m subplot is sampled for presence/absence of plants. At the 10 m by 10 m and 100 m by 100 m scales, only presence and absence of plants were recorded. Each species is recorded only once during sampling, such that an observation of a species at a finer-resolution subplot prevents it from being recorded again if it is encountered in a coarser-resolution subplot. A full dataset for each NEON plant diversity plot was generated by combining all data from all subplots within the 400 m x 400 m boundary, and removing duplicates (which may occur across the 100 m x 100 m subdivisions). 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 on the NEON plant master taxonomy list see NEON.DOC.014042 (https://data.neonscience.org/api/vo/documents/NEON.DOC.014042vK).

**Data Wrangling Decisions** Sampling at the 1 m x 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 x 1 m, 10 m x 10 m, and 100 m x 100 m 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, 100, and 10000). Because of the nested sampling design of the plant data, to capture all records within a subplot at 10 m by 10 m scale, we incorporated all data from both the 1 m by 1 m and 10 m by 10 m scales for that subplot. Similarly, to obtain all records within a subplot at the 100 m by 100 m scale, we included all data from that subplot. 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. Sampling begins on a low intensity schedule with one sampling bout every six weeks. Once >5 ticks of any life stage have been collected within the last year at that site, sampling switches to a high intensity schedule with one bout every three weeks. A site remains on the high intensity schedule until <5 ticks are collected within a year, then it reverts back to the low intensity schedule. High intensity sampling reoccurs when >5 ticks are collected within a year. Onset of 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. Sampling bouts are only initiated if the high temperature on the two consecutive days prior to planned sampling was >0°C.

Ticks are sampled by walking the perimeter of a 40 m x 40 m plot using a 1 m x 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 80-180 meters. 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 and are immediately transferred to a vial containing RNA stabilization solution.

Ticks are sent to the US National Tick Collection at Georgia Southern University for identification to species, life stage, and sex. A subset of nymphal ticks are sent to the Laboratory of Medical Zoology at the University of Massachusetts Amherst for pathogen testing. *Ixodes scapularis* 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). *Amblyomma americanum* was tested for *Anaplasma phagocytophilum*, *Borrelia lonestari* (and other undefined *Borrelia* species), *Ehrlichia chaffeensis*, *Ehrlichia ewingii*, *Francisella tularensis*, and *Rickettsia rickettsii*.

**Data Wrangling Decisions** We downloaded the tick drag and taxonomic identification data (DP1.10093.001; tck_taxonomyProcessed hereafter referred to as 'taxonomy data'; tck_fielddata hereafter referred to as 'field data') via the NEON API. We note that end 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. Some identifications (e.g., IXOSPP) represent potentially mixed samples with more than one species present. Users should exercise caution when computing metrics like richness from these data. Finally, we assigned those samples that were not identified in the lab 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.

First, in the field data, we removed samples that had sampling issues and samples with no count data, which mostly consisted of 2019 data that had not been processed. Next, we removed field samples that did not have corresponding taxonomic information. Conversely, we removed taxonomy samples that did not have corresponding field information, many of which were legacy samples where larvae were not counted. We retained these records where no ticks were found in a sampling bout. In the taxonomy data, we removed samples that did not have an "OK" sample condition.

We left-joined the field data to the taxonomy data to preserve zeros in the count data. To merge the two datasets, we combined 'male' or 'female' into one 'adult' class. Next, we reconciled differences in counts between the two datasets, generally placing greater confidence in the counts from the taxonomy lab (note that beginning in 2019, counts are only taken by taxonomists, so reconciling field and lab counts will no longer be necessary). If the lab determined no ticks were present in the sample (e.g., ticks were mis-identified in the field), we corrected the field data to match the taxonomy data. When there were discrepancies between the identified life-stages but not the total counts of ticks, we retained the counts from the lab. Larvae were not always identified or counted in the lab, or if they were, were only counted up to a limit. If there were more larvae in the field than in the lab (7% of samples), we assigned the remaining unidentified larvae to the order level (IXOSP2). Similarly, in some cases the lab only identified ticks up to a certain invoicing limit, usually 500 individuals. In these cases (2% of samples) we assigned remaining unidentified ticks to the order level.

In cases where counts were off by only a few individuals (3% of samples), we attributed these to miscounts in the field or ticks lost in transit and trusted the taxonomy counts. In cases where one to two ticks in the field data were missing but the discrepancy was >10% of the total count, we assigned those missing ticks to the order level in an adult-specific column. In cases with larger discrepancies without an obvious remark describing it (0.1% of cases), we removed the sample from the dataset, making sure that these cases were <1% of the size of the total dataset. We note that the majority of samples (~85%) had no discrepancies between the lab or field, therefore this process could be ignored by users for whom analyses are not sensitive to exact counts.

For tick pathogens, we downloaded tick pathogen data (DP1.10092.001; tck_pathogen hereafter referred to as 'pathogen data'; tck_pathogenqa hereafter referred to as 'quality data') via the NEON API. 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 and thoroughly test IXOSCA and AMBAME species only. *Borrelia burgdorferi* and *Borrelia burgdorferi sensu lato* tests were merged, since the former was an incomplete pathogen name and likely referred to *B. burgdorferi sensu lato* as opposed to *sensu stricto* (Rudenko et al. 2011). Tick host species was not included in the

original data sheet, but we manually added these to the cleaned data sheets by extracting tick species from subsampleID names.

# 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. Samplers vary by habitat and include Surber, Hess, hand corer, modified kicknet, D-frame sweep, and petite ponar samplers. Lake sampling occurs with a petite ponar near buoy, inlet, and outlet sensors, and D-frame sweeps in littoral zones. Riverine sample collections in deep waters or near 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, but 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** We downloaded the aquatic macroinvertebrate taxonomic identification, data quality, and related field data (DP1.20120.001; Macroinvertebrate collection) using the NEON API. Aquatic macroinvertebrates 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 files, 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/river bottom.

## MicroAlgae (Periphyton and Phytoplankton)

**NEON Sampling Design** Algal sampling methods vary by system (i.e., wadeable streams, rivers, lakes) and are sampled three times per year (i.e., spring, summer, and fall). In wadeable streams, which have variable habitats (e.g., riffles, runs, pools, step pools), five periphyton samples are collected in the dominant habitat type and three in the second most dominant habitat type over a 1km reach. No two samples should come from the sample habitat unit (i.e., same riffle).

Periphyton samples are collected from natural surface substrata with the specific collection method and sampler type dependent on the substrate type (i.e., cobble vs. silt vs. woody debris) (see Moulton II et al. 2002 for detailed methods). Periphyton is also collected from lakes and rivers from five areas in the littoral zone (i.e., shoreline) with the most dominant substratum type selected. In rivers, phytoplankton are sampled near the sensor buoy and at two other deep-water points in the main channel, using a Kemmerer sampler. For lakes, phytoplankton are collected by the central sensor buoy as well as at the inlet and outlet sensor sensors, using a Kemmerer sampler. For stratified lakes and non-wadeable streams, the 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, the phytoplankton sample is a composite from one surface sample, one sample just above the bottom of the euphotic zone, and if the euphotic zone is > 5 m, then one mid-euphotic zone sample.

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. Aliquots for biomass measurements are filtered onto glass-fiber filters and processed for ash-free dry mass.

**Data Wrangling Decisions** We downloaded the algae taxonomic identification, biomass and related field data (DP1.20166.001; 'Periphyton, Seston and Phytoplankton Collection' hereafter referred to as alg_tax_long, alg_biomass and alg_field_data) via the NEON API. 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. 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. Many samples in the alg_tax_long file were missing data in the perBottleSampleVolume field, so a new field called perBSVol was created using NEON domain lab sample volumes. 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 perBSVol. Benthic sample results are expressed in terms of area (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-dried mass (ADFM) 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 NEON-designated ecoregions including both lotic (23 stream habitats) and lentic (five lake habitats) habitats. Each site is surveyed with two bouts per year, during the spring and fall, with a combination of random and

fixed sampling areas (i.e., segments in lakes; reaches in streams) to capture spatial and temporal heterogeneity of aquatic habitats. Each sampling bout is completed within five days 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.

At each lake, 10 sector-shaped segments are established, wherein each segment ranges from the riparian zone into the lake center, therefore effectively capturing both nearshore and offshore habitats. Three of the 10 segments are fixed, which encompasses habitat features most representative of the entire lake, and are surveyed twice a year with a backpack electrofisher using a three-pass electrofishing depletion approach (Moulton II et al. 2002, Peck et al. 2006). All three passes in a fixed sampling segment or reach are completed on the same day, with a minimum gap of 30-minutes between successive passes. Additionally, a fyke net and a gill net are deployed at each fixed segment (Baker et al. 1997). The remaining random segments are sampled on a rotation design with a single electrofishing pass, one mini-fyke net, and one gill net where three segments are surveyed twice a year.

In each stream site, a maximum of 10 non-overlapping reaches (each reach 70-130m in length) are designated within a 1km stream length, with six of those reaches sampled during any given sampling bout. The 10 reaches include three fixed reaches that are consistently sampled and seven random reaches that are sampled on rotation. The three fixed reaches encompass all representative habitats found within a 1-km-stretch and are surveyed twice a year (once in each bout) using a three-pass electrofishing depletion approach (Moulton II et al. 2002, Peck et al. 2006). The random reaches are surveyed on rotation via a single-pass depletion approach; three random reaches are sampled twice a year (once in each bout), while another three random reaches are sampled in the subsequent year. Electrofishing at streams is done during daytime. At lake sites, electrofishing is started and ceases 30-minutes after and before sunset and sunrise, respectively, with a maximum of five passes per sampling bout. Gill nets are deployed for 1-2 hrs either in the morning or early afternoon, while fyke nets are positioned before sunset and recovered after sunrise on the following day. Precise start and end times for both electrofishing and net deployments are documented.

In all surveys, captured fish are identified to the lowest practical taxonomic level, and morphometrics (i.e., body mass and body length) are recorded before releasing. Relative abundance for each fish taxon is also recorded by direct enumeration or estimation (i.e., by scooping and counting the total number of specimens in one 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 (up to 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, photographed, 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 (1) either identified to the finest possible taxonomic resolution or 2) 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). Local fish taxonomists in each NEON domain identify the morphospecies and species with uncertain taxonomic identities.

Fish data (taxonomic identification and relative abundance) are recorded per each sampling reach in streams or segment in lakes in each bout. 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. A reachID column is also provided that 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.

We downloaded all fish data (i.e., fsh_perPass, fsh_fieldData, fsh_bulkCount, fsh_perFish), including the complete taxon table for fish, for both stream and lake sites surveyed via the NEON API. We joined the fsh_perPass, fsh_fieldData, and fsh_bulkCount datasets to produce a table with bulk-processed data that merged fsh_perPass, fsh_fieldData, and fsh_perFish to

concatenate individual-level data. Finally both individual-level and bulk-processed datasets were appended into a single table. 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). In this case, 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 7 NEON lake sites across 4 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 location shallower than 4 m. This results in 3 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** We downloaded the zooplankton taxonomic identification and related field data (DP1.20219.001). 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. 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

Table 1: **Summary of data products included in this study.**

| taxa | data product ID | data_product | n_site | n_species | start_year | end_year | modify_time |
|---|---|---|---|---|---|---|---|
| algae | DP1.20166.001 | data_algae | 33 | 1824 | 2014 | 2019 | 2020-10-30 |
| beetle | DP1.10022.001 | data_beetle | 47 | 756 | 2013 | 2020 | 2020-11-10 |
| bird | DP1.10003.001 | data_bird | 47 | 535 | 2013 | 2019 | 2020-12-08 |
| fish | DP1.20107.001 | data_fish | 27 | 125 | 2016 | 2020 | 2020-11-11 |
| herp_bycatch | DP1.10022.001 | data_herp_bycatch | 41 | 125 | 2014 | 2020 | 2021-01-03 |
| macroinvertebrate | DP1.20120.001 | data_macroinvertebrate | 34 | 1276 | 2014 | 2020 | 2020-10-30 |
| mosquito | DP1.10043.001 | data_mosquito | 47 | 126 | 2015 | 2020 | 2020-10-30 |
| plant | DP1.10058.001 | data_plant | 47 | 6075 | 2013 | 2020 | 2020-12-09 |
| small_mammal | DP1.10072.001 | data_small_mammal | 46 | 137 | 2014 | 2019 | 2020-12-08 |
| tick | DP1.10093.001 | data_tick | 41 | 19 | 2014 | 2018 | 2020-10-30 |
| tick_pathogen | DP1.10092.001 | data_tick_pathogen | 14 | 12 | 2013 | 2020 | 2020-10-30 |
| zooplankton | DP1.20219.001 | data_zooplankton | 7 | 154 | 2014 | 2020 | 2020-12-16 |

count data to "count per liter."

# Results (or how to get and use tidy NEON organismal data)

All cleaned data products can be obtained from the R package neonDivData, which can be installed from Github. Installation instructions can be found on the Github webpage (https://github.com/daijiang/neonDivData). Table 1 shows the brief summary of all data products. To get a specific data product, we can just call the objects in the data_product column in Table 1. 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 and species measurements were provided by NEON for some taxonomic groups, they are also included in these data products. Information such as latitude, longitude, and elevation for all taxonomic groups were saved in the neon_locations object of the R package, which is equivalent to the "sampling_location" table of the ecocomDP data format. Information about species scientific names and identification references 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 products, 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 below.

```
library(dplyr)
library(neonDivData)
# get species richness at each site
sp_rich_plant = data_plant %>%
  group_by(siteID) %>%
  summarise(nspp = n_distinct(taxonID))
# get latitude and longitude of all sites
sp_rich_plant = left_join(sp_rich_plant,
                          filter(neon_locations, taxa == "plant") %>%
                            # each site has multiple plots with slightly different lat/long
                            group_by(siteID) %>%
                            summarise(decimalLatitude = mean(decimalLatitude),
                                      decimalLongitude = mean(decimalLongitude)))
```

Figure 2 shows the utility of NEON data for exploring macroecological patterns. One of the most well known and studied macroecological patterns is the latitudinal biodiversity gradient, wherein sites are more speciose at lower latitudes relative to higher latitudes (Fischer 1960, Hillebrand 2004). Herbaceous plants of NEON generally follow this pattern, but interestingly, Read et al. (2018) found that NEON small mammal data do not support this pattern.

In addition to allowing for quick exploration of macroecological patterns of diversity, the data products presented in this paper also enable investigation of effects of taxonomic resolution on diversity indices since taxonomic information is preserved for observations under family level for all taxonomic 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 having greater uncertainty. 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 product 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

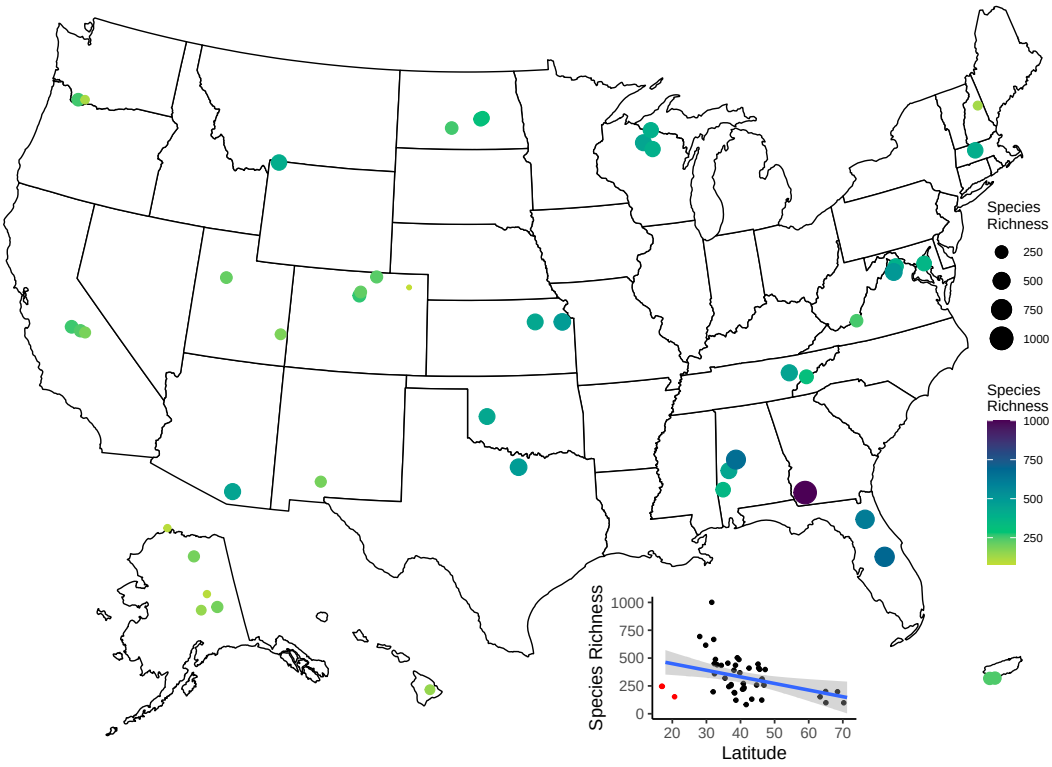


Figure 2: Plant species richness mapped across NEON terrestrial sites. The inset scatterplot shows latitude on the x-axis and species richness on the y-axis, with red points representing sites in Puerto Rico and Hawaii. Alaska, Hawaii, and Puerto Rico were rearranged to save space.

site for data subsetted at the genus, species, and subspecies level. Jost indices are essentially Hill Numbers quantifying species diversity 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) indicate 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 NEON beetle data streams 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.

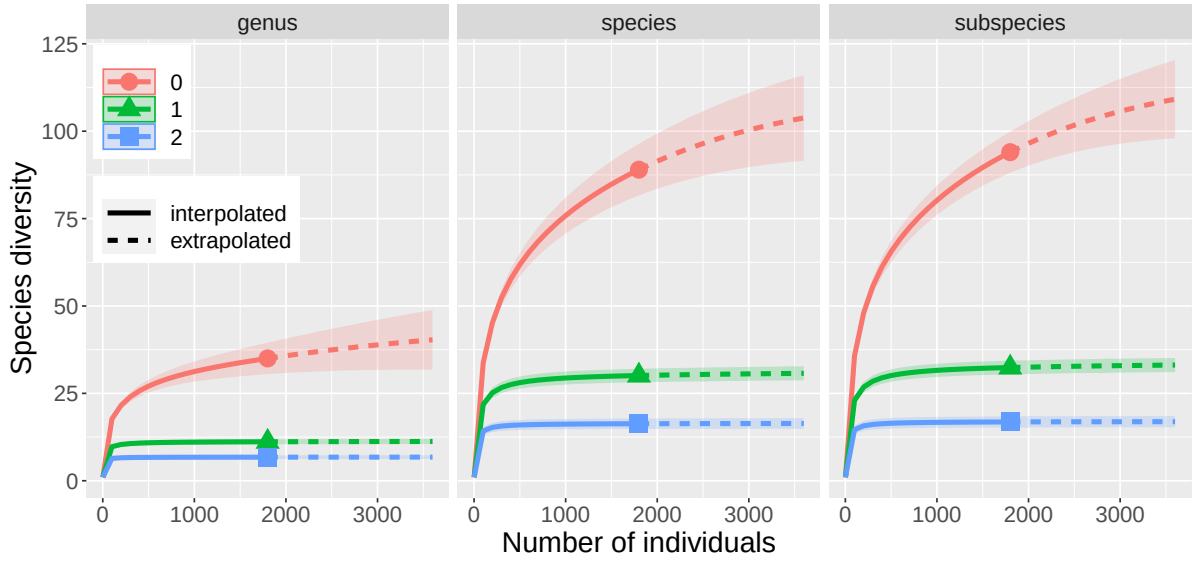


Figure 3: Rarefaction of beetle abundance data from collections made at the Oak Ridge National Laboratory (ORNL) National Ecological Observatory Network (NEON) site from 2014-2020 generated using the iNEXT package in R (Hsieh et al. 2016) based on different levels of taxonomic resolution (i.e., genus,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

# Discussion (or how to maintain and update tidy NEON organismal data)

NEON organismal data hold lots of potential to understand biodiversity change across space and time Jones et al. (2021). Multiple biodiversity research and education programs have used NEON data even before NEON became fully operational in May 2019 [CITATION]. With the expected large investment to maintain NEON over the next 30 years, NEON organismal data, alone or coupled with other major environmental datasets, will be invaluable to help us understand and track biodiversity change in an era of fast environmental 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). Whereas the data collection protocols implemented by NEON staff are reproducible across space and time, the studies NEON data users perform after downloading NEON's open data will only be fully repeatable if the steps made to generate and analyze derived data sets are documented (Ellison 2010). If such provenance is not curated for studies using NEON data, then 30 years from now we risk a situation where we find ourselves puzzling over inconclusive results from studies that use the same data. However, with careful provenance the gains in macroscale environmental knowledge have the potential to be transformative. By providing a standardized and easy-to-use data product 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 data are reproducible.

Unlike NEON's organismal data, most long-term ecological community observations are made up of data where sampling protocols vary. The ecocomDP data design pattern was recently created to harmonize such community organismal observations to promote biodiversity synthesis efforts [O'Brien et al. In review; Record et al. (2020)]. The ecocomDP harmonized format is not bound by any research question (so long as it pertains to community level data), adds relevant metadata on the spatial and temporal scale of the derived data, and is amenable to different types of measurements (e.g., count, percent cover, biomass). The data package we

present in this paper converts NEON organismal data to the ecocomDP format core tables (i.e., observation, taxon, location), so that NEON data will be more easily integrated with other community data sets, especially those available in the EDI repository (e.g., LTER, LTREB data). In the EDI repository, derived data sets are given DOIs, so that differences may be tracked from original to derived data sets. The data presented in this paper have their own DOIs that are discoverable through EDI and may be cited.

There are some important notes about the data product we provided. First, we did not check the taxonomy of all groups given that NEON already did its best to make sure that species identifications are correct. However, not every record was identified to species, with genus, family, or even order level species IDs in all groups. IDs above genus level may not be useful for most biodiversity projects. We thus decided to remove records with such IDs for groups that are relatively easier to identify (fish, plant, small mammals) or have very few taxon IDs that are above genus level (mosquito). However, for groups that are hard to identify (algae, beetle, bird, macroinvertebrate, tick, and tick pathogen), we decided to keep all records no matter what level of taxon IDs they have. 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 think carefully about which level of taxon IDs they need for their research. Second, we kept records without any observed species/individuals in some taxonomic groups (beetle, bird, small mammal). Such empty records still provide information about sampling efforts, which can be critical for some projects to control for. For example, if two sites both have the same number of small mammals during the same time period, however, one site has many more empty records (traps), then we can infer that the abundances of small mammals are lower than the other side. If such information is not needed in a project, then we can simply remove them. For algae and birds, we added standardized measurements as density and number of fish caught per hour, respectively, in the data product. Third, there are other organismal groups (Aquatic Microbes and Aquatic Plants; soil microbes were covered in Qin et al. in this issue) not included in this study given the complexity of microbial data.

All codes that conducted the Data Wrangling Decisions were available online (Github and Zenodo, URL HERE). Therefore, users can easily reproduce the standardized data product, modify the code if they need to make different decisions during the data wrangling process, and correct

any mistakes of our code by submitting a pull request to our repository. It is also easy to update the standardized data product when new data is uploaded by NEON to their data portal because the whole data wrangling workflow was automated. In fact, our Github repository is scheduled to run the whole workflow every year.

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, upload the data to the EDI repository, and cite their data with the discoverable DOI given to them by EDI. Note that ecocomDP data design package was intended for community ecology analyses, so it will not handle individual data, such as those required for demographic analyses, well. A promising package for demographic analyses of NEON data is popler (Compagnoni et al. 2020). The notion of different data packages for different levels of biological organization and different types of data raises the question of whether there are other data design packages that the NEON user community should adopt at the inception of the observatory. As the observatory moves forward, this is an important discussion to be had by 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). It will also be important moving forward to make sure that funding for the maintenance of repositories, such as EDI, remains to prevent the loss of archived derived data sets.

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 consisting 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, coming to many of these Data Wrangling Decisions benefited greatly from 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 in terms of transformative scientific discoveries.

# Conclusion

Macrosystems ecology 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 strive to make our analyses and all derived data as reproducible as possible to ensure that connections across studies will be possible. Harmonized data sets will help in this endeavour because they naturally promote the collection of provenance as data are collated into derived products (O'Brien et al. In review, Reichman et al. 2011). Harmonized data also make synthesis easier because efforts to clean and format data leading up to analyses do not have to be repeatedly performed by individual researchers (O' Brien et al. In review). The derived data products presented here in the ecocomDP format illustrate a potential path forward in achieving a reproducible framework for data derived from NEON organismal data for community ecology analyses. *Highlight value of collaboration between NEON user community and NEON staff for advancing NEON enabled science.*

# 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. 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.

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.

Compagnoni, A., A. J. Bibian, B. M. Ochocki, S. Levin, K. Zhu, and T. E. Miller. 2020. Popler: An r package for extraction and synthesis of population time series from the long-term ecological research (LTER) network. Methods in Ecology and Evolution 11:258–264.

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.

Ellison, A. M. 2010. Repeatability and transparency in ecological research. Ecology 91:2536–2539.

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.

Fischer, A. G. 1960. Latitudinal variations in organic diversity. Evolution 14:64–81.

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.

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.

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.

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.

Hoekman, D., Y. P. Springer, C. Barker, R. Barrera, M. Blackmore, W. Bradshaw, D. H. Foley, H. S. Ginsberg, M. Hayden, C. Holzapfel, and others. 2016. Design for mosquito abundance, diversity, and phenology sampling within the national ecological observatory network. Ecosphere 7:e01320.

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.

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.

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.

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.

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.

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.

915 Thorpe, A. S., D. T. Barnett, S. C. Elmendorf, E.-L. S. Hinckley, D. Hoekman, K. D. Jones, K. E.
916 LeVan, C. L. Meier, L. F. Stanish, and K. M. Thibault. 2016. Introduction to the sampling
917 designs of the n ational e cological o bservatory n etwork t errestrial o bservation s ystem.
918 Ecosphere 7:e01627.

919 Vellend, M., L. Baeten, I. H. Myers-Smith, S. C. Elmendorf, R. Beauséjour, C. D. Brown, P. De
920 Frenne, K. Verheyen, and S. Wipf. 2013. Global meta-analysis reveals no net change in
921 local-scale plant biodiversity over time. Proceedings of the National Academy of Sciences
922 110:19456–19459.

923 Wilkinson, M. D., M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg,
924 J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, and others. 2016. The FAIR guiding principles
925 for scientific data management and stewardship. Scientific data 3:1–9.

926 Worm, B., and D. P. Tittensor. 2018. A theory of global biodiversity (MPB-60). Princeton
927 University Press.