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

16 December, 2020

Abstract: Authors of this paper are all interested in using NEON data for biodiversity research. We have spent lots of time reading the documentations and cleaning up the data for our own studies. We believe that we can document our data cleaning process and provide the tidy NEON data for the community so that others can use the data readily for biodiversity research.

Key words: NEON, Biodiversity, Data

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, 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 is currently limited by the lack of standardized high quality and open-access data across large spatial scales and long time periods.

There is currently a major effort underway to overcome the issues above. For example, the Long Term Ecological Research Network (LTER) consists of 28 sites that provide long term datasets for a diverse set of ecosystems. However, there is no standardization in the design and data

collections across LTER sites. The National Ecological Observatory Network (NEON) is a continental-scale observatory network that collects long-term, standardized, and open access datasets broadly aimed at enabling better understanding of how U.S. ecosystems change through time (Keller et al. 2008, Thorpe et al. 2016). Data collected 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 will 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 principles (findable, accessible, interoperable, and reusable Wilkinson et al. 2016), but there are still hurdles to overcome in the use of their organismal data for biodiversity research. For example, different NEON data products use different field names for similar measurements, some include sampling unit information while this must be calculated for others, etc. Furthermore, NEON organismal data products provide lots of raw data, most of them may not be needed 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 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 get to that standard data format is different and not always obvious for each NEON organismal data product. 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.

Here, our goal is to provide a standardized "tidy version" of NEON organismal datasets 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 large ecology community, and to facilitate the use of NEON data to advance biodiversity research.

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

General points we could go over that apply to all data sets before going into details:

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 et al. 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, 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.).

Ground Beetles

NEON Sampling Design Each site is sampled via pitfall trap, with 10 separate distributed plots at each site and four pitfall traps at each plot initially - placed in the ground at the cardinal direction points 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 are sorted and identified by NEON technicians, after which a subset of individuals are sent to be pinned and re-identified by an expert taxonomist. More details can be found in Hoekman et al.

(2017).

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 was 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.

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 in order 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.

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 light/fan assembly (battery powered), 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 the trap is collected.

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 a 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, only if temperatures are above 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 8 hours after trap is set, and one day-active sample, taken at dusk or about 16 hours after the trap is set. Thus, a 24-hour sampling bout yields 20 "samples" from ten 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 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, collect date, and sample or subsample ID for all data frames. We then merged all four data frames into one while checked carefully during the process. In the merged data frame, we only kept records that have target taxa (i.e., targetTaxaPresent == "Y") or have no known compromised sampling condition (i.e., sampleCondition == "No known compromise"). We further removed a small number of records with species identified at 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 individualCount * (totalWeight / 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 g (Thibault et al. 2019). In North America, this includes cricetids, heteromyids, small sciurids, and introduced murids, but excludes shrews, large squirrels, rabbits, or weasels, despite the fact that individuals of these species may be incidentally captured. A total of 65 species across the US meet these criteria and are designated as "target species' '. 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 ten rows and ten 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. 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). Records were also removed 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 m^2 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 total 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 sampled during one or two bouts per year, and are a total of 400 m x 400 m. Sampling is done using a nested design, where the entire plot is first subdivided into 4 100 m x 100 m subplots. For each of these, one or more 1 m x 1 m nested subplots are then sampled; species coverages within the 1 m^2 area were estimated visually. Next, one or more 10 m x 10 m subplots, inside of which the finer resolution subplots are located, are sampled. Finally, the 100 m x 100 m subplot is sampled. At 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 fine-resolution subplot prevents it from being recorded again if it is encountered in a coarser scale 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 in 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.

Data Wrangling Decisions Because 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, etc), we removed records with divDataType as "otherVariables". We also removed records whose targetTaxaPresent is N (i.e., non a target species). Additionally, for all spatial resolution with observatory data (i.e., 1 m x 1 m, 10 m x 10 m, and 100 m x 100 m data), any record lacking information critical to 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 morphspecies) were not considered. To stack data from different spatial resolution into one data frame, we created a pivot column named as sample_area_m2 (possible values are 1, 100, and 10000). Because of the nested design of plant data, to get all records within a subplot at 10 m by 10 m scale, we need to use all data from both 1 m by 1 m and 10 m by 10 m scales for that subplot; similarly, to get all records within a subplot at 100 m by 100 m scale, we need to include 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

NEON Sampling Design

Data Wrangling Decisions

Tick pathogens

NEON Sampling Design

Data Wrangling Decisions

Aquatic Organisms

MicroAlgae (Periphyton and Phytoplankton)

NEON Sampling Design

Data Wrangling Decisions

Fish

NEON Sampling Design

Data Wrangling Decisions

Aquatic macroinvertebrates

NEON Sampling Design

Data Wrangling Decisions

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

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

taxa	neon_DPI	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	2020-11-23
fish	DP1.20107.001	data_fish	27	125	2016	2020	2020-11-11
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	2020-10-30
small_mammal	DP1.10072.001	data_small_mammal	46	145	2014	2019	2020-10-30
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

(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 x. Such data products include cleaned (and standardized if needed) occurrence data for the taxonomic groups covered. 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. Information about species scientific names and identification references of all taxonomic groups were saved in the neon_taxa object.

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. 1). To show how easy it is to get site level species richness, we presented the code used to generate the data for Fig. 1 below.

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 (Balch et al. 2019). 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. By providing a standardized

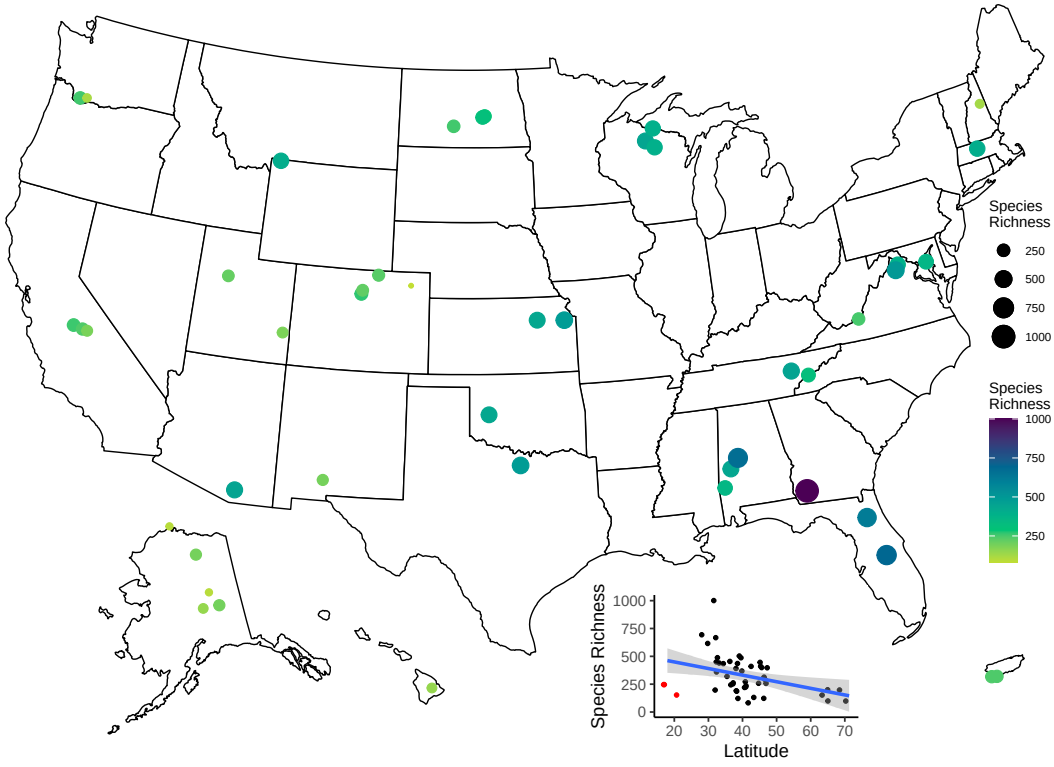

Figure 1: Distribution of plant species richness across all NEON terrestrial sites. Alaska, Hawaii, and Puerto Rico were rearranged to save space.

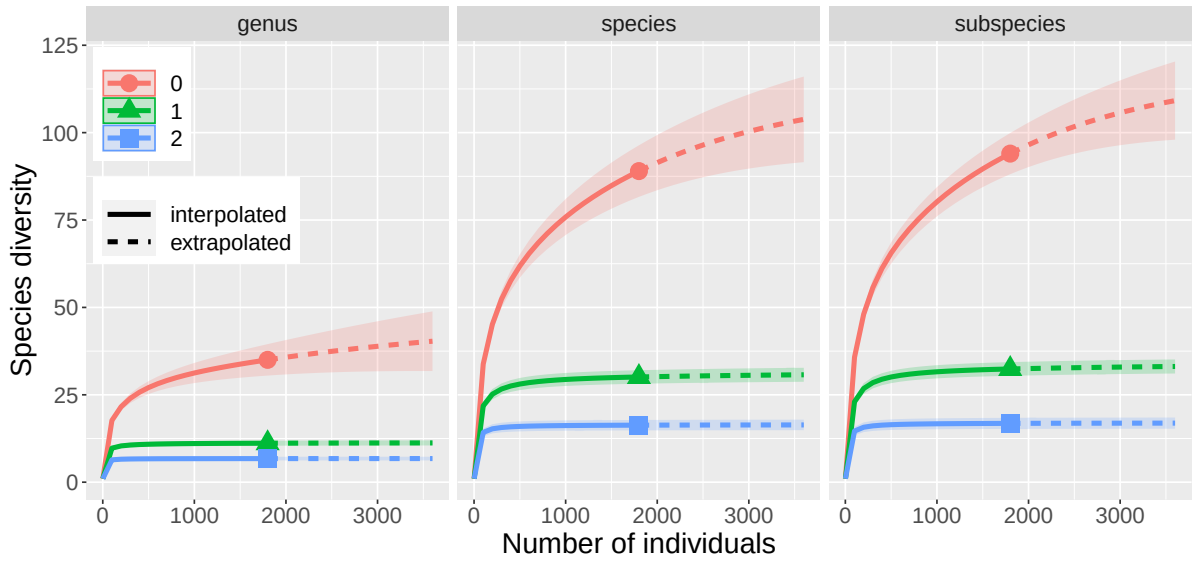

Figure 2: Rarefaction of beetle abundance data from collections made at the Oak Ridge National Laboratory (ORNL) National Ecological Observatory Network (NEON) site from 2014-2020 generated using the iNEXT package in R (Hsieh et al. 2016) based on different levels of taxonomic resolution (i.e., genus, species, subspecies). Different colors indicate Jost Indices (i.e., Hill Numbers quantifying species diversity that vary in how abundance is weighted with the 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 (q = 1; Jost 2006).

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.

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. 2). 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.

- All code are open and freely available
- Plans for future maintenance and (automated) updates.

Conclusion

highlight value of harmonized data for community of researchers to advance the field. Highlight value of collaboration between NEON user community and NEON staff for advancing NEON enabled science.

Reference

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.

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.

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

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.

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.

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.

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.

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.

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.

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.