Standardized NEON organismal data for biodiversity research

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

13 April, 2022

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

† Equal contributions

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

Open Research Statement

No data were collected for this study. All original data were collected by NEON and are publicly available at NEON's data portal. We standardized such data and provided them as a data package, which is available at Github (https://github.com/daijiang/neonDivData) and archived at Zenodo (https://doi.org/10.5281/zenodo.6419751). Data were also permanently archived at the EDI data repository (Li et al. 2022) (https://doi.org/10.6073/pasta/c28dd4f6e7989003505ea02e9a92afbf).

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

examples of how the standardized data may be used for biodiversity research. By providing a standardized data package, we hope to enhance the utility of NEON organismal data in advancing biodiversity research and encourage the use of the harmonized ecocomDP data design pattern for community ecology data from other ecological observatory networks.

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

Introduction (or why standardized NEON organismal data)

A central goal of ecology is to understand the patterns and processes of biodiversity, and this is particularly important in an era of rapid global environmental change (Midgley and Thuiller 2005, Blowes et al. 2019). Such understanding is only possible through studies that address questions like: How is biodiversity distributed across large spatial scales, ranging from ecoregions to continents? What mechanisms drive spatial patterns of biodiversity? Are spatial patterns of biodiversity similar among different taxonomic groups, and if not, why do we see variation? How does community composition vary across spatial and environmental gradients? What are the local and landscape scale drivers of community structure? How and why do biodiversity patterns change over time? Answers to such questions will enable better management and conservation of biodiversity and ecosystem services.

Biodiversity research has a long history (Worm and Tittensor 2018), beginning with major scientific expeditions (e.g., Alexander von Humboldt, Charles Darwin) aiming to document global species lists after the establishment of Linnaeus's Systema Naturae (Linnaeus 1758). Beginning in the 1950's (Curtis 1959, Hutchinson 1959), researchers moved beyond documentation to focus on quantifying patterns of species diversity and describing mechanisms underlying their heterogeneity. Since the beginning of this line of research major theoretical breakthroughs (MacArthur and Wilson 1967, Hubbell 2001, Brown et al. 2004, Harte 2011) have advanced our understanding of potential mechanisms causing and maintaining biodiversity. Modern empirical studies, however, have been largely constrained to local or regional scales and focused on one or a few taxonomic groups, because of the considerable effort required to collect observational data. There are now unprecedented numbers of observations from independent

small and short-term ecological studies. These data support research into generalities through syntheses and meta-analyses (Vellend et al. 2013, Blowes et al. 2019, Li et al. 2020), but this work is challenged by the difficulty of integrating data from different studies and with varying limitations. Such limitations include: differing collection methods (methodological uncertainties); varying levels of statistical robustness; inconsistent handling of missing data; spatial bias; publication bias; and design flaws (Martin et al. 2012, Nakagawa and Santos 2012, Koricheva and Gurevitch 2014, Welti et al. 2021). Additionally, it has historically been challenging for researchers to obtain and collate data from a diversity of sources for use in syntheses and/or meta-analyses (Gurevitch and Hedges 1999).

Barriers to meta-analyses have been reduced in recent years to bring biodiversity research into the big data era (Hampton et al. 2013, Farley et al. 2018) by large efforts to digitize museum and herbarium specimens (e.g., iDigBio), successful community science programs (e.g., iNaturalist, eBird), technological advances (e.g., remote sensing, automated acoustic recorders), and long running coordinated research networks. Yet, each of these remedies comes with its own limitations. For instance, museum/herbarium specimens and community science records are increasingly available, but are still incidental and unstructured in terms of the sampling design, and exhibit marked geographic and taxonomic biases (Martin et al. 2012, Beck et al. 2014, Geldmann et al. 2016). Remote sensing approaches may cover large spatial scales, but may also be of low spatial resolution and unable to reliably penetrate vegetation canopy (Palumbo et al. 2017, G Pricope et al. 2019). The standardized observational sampling of woody trees by the United States Forest Service's Forest Inventory and Analysis and of birds by the United States Geological Survey's Breeding Bird Survey have been ongoing across the United States since 2001 and 1966, respectively (Bechtold and Patterson 2005, Sauer et al. 2017), but cover few taxonomic groups. The Long Term Ecological Research Network (LTER) and Critical Zone Observatory (CZO) both are hypotheses-driven research efforts built on decades of previous work (Jones et al. 2021). While both provide considerable observational and experimental datasets for diverse ecosystems and taxa, their sampling and dataset design are tailored to their specific research questions and a priori, standardization is not possible. Thus, despite recent advances biodiversity research is still impeded by a lack of standardized, high quality, and open-access data spanning large spatial scales and long time periods.

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

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

One attractive standardized formatting style for NEON organismal data is that of ecocomDP (ecological community data design pattern, O'Brien et al. 2021). EcocomDP is the brainchild of members of the LTER network, the Environmental Data Initiative (EDI), and NEON staff, and provides a model by which data from a variety of sources may be easily transformed into consistently formatted, analysis ready community-level organismal data packages. This is done

using reproducible code that maintains dataset "levels": L0 is incoming data, L1 represents an ecocomDP data format and includes tables representing observations, sampling locations, and taxonomic information (at a minimum), and L2 is an output format. Thus far, >70 LTER organismal datasets have been harmonized to the L1 ecocomDP format through the R package ecocomDP and more datasets are in the queue for processing into the ecocomDP format by EDI (O'Brien et al. 2021).

We standardized NEON organismal data into the ecocomDP format and all R code to process NEON data products can be obtained through the R package ecocomDP. For the major taxonomic groups included in this initial effort, NEON sampling designs and major data wrangling decisions are summarized in the Materials and Methods section. We archived the standardized data in the EDI Data Repository. To facilitate the usage of the standardized datasets, we also developed an R data package, neonDivData. We refer to the input data streams provided by NEON as data products, whereas the cleaned and standardized collection of data files provided here as objects within the R data package, neonDivData, across this paper. Standardized datasets will be maintained and updated as new data become available from the NEON portal. We hope this effort will substantially reduce data processing times for NEON data users and greatly facilitate the use of NEON organismal data to advance our understanding of Earth's biodiversity.

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

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

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

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

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