Received: 4 March 2024 | Revised: 17 July 2024 | Accepted: 2 August 2024

DOI: 10.1002/ecs2.70159

ECOSPHERE
AN ESA OPEN ACCESS JOURNAL

SYNTHESIS & INTEGRATION

Special Feature: Harnessing the NEON Data Revolution

Recommendations for developing, documenting, and distributing data products derived from NEON data

Jeff W. Atkins[1,2] | **Kelly S. Aho**[3,4] | **Xuan Chen**[5] | **Andrew J. Elmore**[6] | **Rich Fiorella**[7] | **Wenqi Luo**[8] | **Danica Lombardozzi**[9,10] | **Claire Lunch**[11] | **Leah Manak**[12] | **Luis X. de Pablo**[13] | **Allison N. Myers-Pigg**[14] | **Sydne Record**[15] | **Tong Qiu**[16] | **Samuel Reed**[17] | **Benjamin Ruddell**[12] | **Brandon Strange**[12] | **Christa L. Torrens**[18] | **Kelsey Yule**[19] | **Andrew D. Richardson**[12,20]

[1]USDA Forest Service, Southern Research Station, New Ellenton, South Carolina, USA

[2]Department of Biology, Virginia Commonwealth University, Richmond, Virginia, USA

[3]Department of Earth & Environmental Sciences, Michigan State University, East Lansing, Michigan, USA

[4]Department of Integrative Biology, Michigan State University, East Lansing, Michigan, USA

[5]Department of Biological Sciences, Salisbury University, Salisbury, Maryland, USA

[6]Appalachian Laboratory, University of Maryland Center for Environmental Science, Frostburg, Maryland, USA

[7]Los Alamos National Laboratory, Los Alamos, New Mexico, USA

[8]School of Environment and Sustainability, University of Michigan, Ann Arbor, Michigan, USA

[9]National Center for Atmospheric Research, Boulder, Colorado, USA

[10]Department of Ecosystem Science and Sustainability, Colorado State University, Fort Collins, Colorado, USA

[11]National Ecological Observatory Network, Battelle, Boulder, Colorado, USA

[12]School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, Arizona, USA

[13]Biofrontiers Institute, University of Colorado, Boulder, Colorado, USA

[14]Pacific Northwest National Laboratory, Marine and Coastal Research Laboratory, Sequim, Washington, USA

[15]Department of Wildlife, Fisheries, and Conservation Biology, University of Maine, Orono, Maine, USA

[16]Nicholas School of the Environment, Duke University, Durham, North Carolina, USA

[17]Natural Resources and Management, University of Minnesota, St. Paul, Minnesota, USA

[18]Flathead Lake Biological Station, University of Montana, Polson, Montana, USA

[19]School of Life Sciences, Arizona State University, Tempe, Arizona, USA

[20]Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, Arizona, USA

Correspondence
Jeff W. Atkins
Email: jwatkins6@vcu.edu

Funding information
National Science Foundation (NSF), Grant/Award Numbers: 2301322, 2242803; Directorate for Biological

Abstract

The National Ecological Observatory Network (NEON) provides over 180 distinct data products from 81 sites (47 terrestrial and 34 freshwater aquatic sites) within the United States and Puerto Rico. These data products include both field and remote sensing data collected using standardized protocols and

Sciences, Grant/Award Number: 2105828; Battelle

Handling Editor: Kristofer D. Johnson

sampling schema, with centralized quality assurance and quality control (QA/QC) provided by NEON staff. Such breadth of data creates opportunities for the research community to extend basic and applied research while also extending the impact and reach of NEON data through the creation of derived data products—higher level data products derived by the user community from NEON data. Derived data products are curated, documented, reproducibly-generated datasets created by applying various processing steps to one or more lower level data products—including interpolation, extrapolation, integration, statistical analysis, modeling, or transformations. Derived data products directly benefit the research community and increase the impact of NEON data by broadening the size and diversity of the user base, decreasing the time and effort needed for working with NEON data, providing primary research foci through the development via the derivation process, and helping users address multidisciplinary questions. Creating derived data products also promotes personal career advancement to those involved through publications, citations, and future grant proposals. However, the creation of derived data products is a nontrivial task. Here we provide an overview of the process of creating derived data products while outlining the advantages, challenges, and major considerations.

KEYWORDS
community science, data, derived data products, NEON, observatory science, Special Feature: Harnessing the NEON Data Revolution

INTRODUCTION

The abundance and accessibility of large, diverse, publicly available data create opportunities for synthesis work across a broad array of disciplines through the creation of higher level derived data products. Derived data products are well-documented, higher level, reproducibly generated synthetic output datasets created by the application of standardized processing steps that leverage one or more existing datasets as inputs. Well-designed and well-documented derived data products are likely to have a broad audience and may see substantial reuse by many researchers, thus benefiting the broader community as well as the creators of the data products—either through direct credit via citations to the product's digital object identifier (DOI), supporting manuscripts, and later manuscripts that use those data products or via bolstering the reputation of the creators (Colavizza et al., 2020).

Derived data products can be generated from all manner of input data that are not publicly available, although data license terms may preclude remixing and redistribution. Thus, standardized, open source data sets are likely to be the most useful to the community and the easiest data to work with. Large networks such as the Long-Term Ecological Research Network (LTER), Long-Term Agricultural Research Network (LTAR), AmeriFlux, FLUXNET, or Integrated Carbon Observation System (ICOS) or centralized data providers like NASA and USGS provide abundant data with liberal reuse policies. Despite such data proliferation, the "right" data for a specific question may not be readily available. Often data are in more of a "raw" form—quality checked and controlled but subjected to minimal processing. Thus, opportunities are abundant for the creation of derived data products to address more specific questions or to facilitate analyses. Here we describe the process of creating derived data products from National Ecological Observatory Network (NEON) data as NEON provides an abundant source of available ecological and environmental data for the creation of derived data products. However, the processes we outline, as well as the considerations and challenges, are applicable to any openly distributed data from any public data source.

NEON is a continental-scale observation facility, sponsored by the National Science Foundation (NSF) and operated under cooperative agreement by Battelle, with the goal to collect long-term, open access ecological data to better understand how ecosystems are changing at continental scales (Keller et al., 2008). NEON represents an exciting frontier, allowing the study of patterns and

processes linking land use and climate change to ecosystem and organismal responses (Heffernan et al., 2014; Peters et al., 2008). NEON provides over 180 distinct data products from 81 sites within the United States and Puerto Rico. These data products include both field and remote sensing data collected using standardized protocols and sampling schema across all NEON sites (Barnett, Adler, et al., 2019; Barnett, Duffy, et al., 2019; Meier et al., 2023; Metzger et al., 2019; Parker & Utz, 2022), with centralized quality assurance and quality control (QA/QC) validation provided by NEON staff. NEON data collection is planned to continue for 30 years with all data freely and openly available for use. The impact of NEON is already apparent. Since becoming fully operational in 2019, NEON has published and updated over 180 data products, with 161 citable via DOIs covering the major themes of the atmosphere, biogeochemistry, ecohydrology, land cover and processes, and organisms, populations, and communities as of January 2024 (DataCite Commons, 2023). Also, as of 2024, over 1000 publications using NEON data, samples, or other assets, with over 18,000 cumulative citations (NEON Dimensions, 2023; GBIF) and 366 grants totaling over US $250M linked to or explicitly connected to NEON (Thibault et al., 2023) (Figure 1).

One major obstacle for many current and potential users of NEON data is that despite the vast abundance of data products provided by NEON, the exact data product needed for a certain analysis may not exist—even if the data needed to produce this product might. For example, if the research goal was to determine differences in the effects of drought conditions on plant productivity among NEON sites, first it would be necessary to calculate the rates of productivity from multiple NEON data sources, and then to calculate drought metrics from precipitation, soil moisture, air temperature, and/or vapor pressure deficit measurements. Each of these efforts is nontrivial and could represent research advances in their own right. Thus, the need for derived data products is evident and creates the opportunity for community engagement to develop products from existing NEON data for community use. Derived data products are provided, produced,

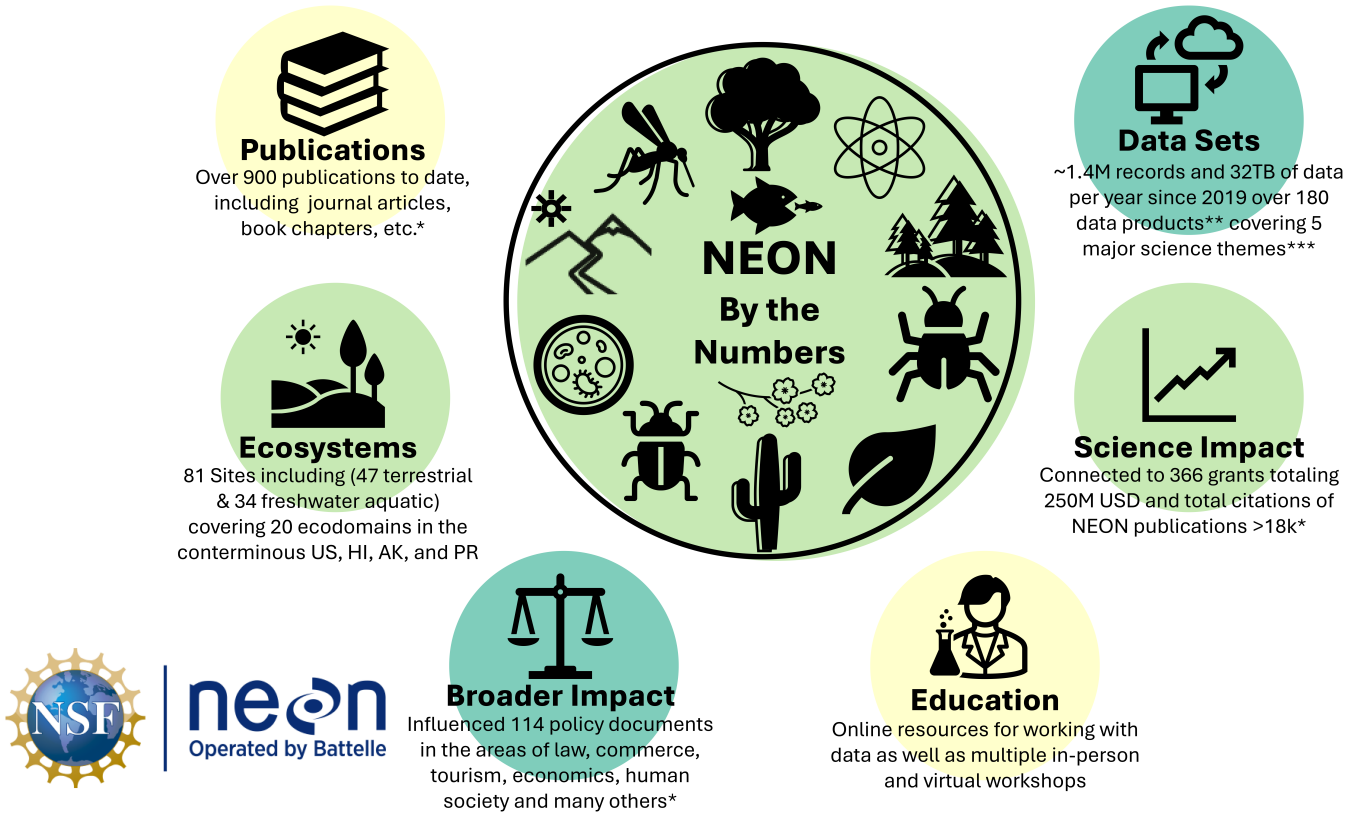

FIGURE 1 An infographic showing National Ecological Observatory Network (NEON) "by the numbers" including details on publications, citations, funding, and the NEON data universe. Data and statistics from NEON Dimensions (2023), Global Biodiversity Information Facility (GBIF), and DataCite Commons (2023) and are current through January 2024. *NEON Dimensions data: https://neon.dimensions.ai/discover/publication. **DataCite Commons data: https://commons.datacite.org/repositories/x3ockqg and https://doi.org/10.5141/jee.23.076. ***Themes: Atmosphere; Biogeochemistry; Ecohydrology; Land Cover & Process; Organisms, Populations, and Communities.

and curated by the community and hosted in repositories other than NEON (e.g., EDI, ESS-DIVE, figshare, Dryad, ORNL DAAC). Derived data products meet specific user community needs, increase the density and diversity of available data products, and ultimately reduce the time and effort required of new users to engage in impactful research at continental scales. However, creating derived data products is not a trivial task. The objective of this paper is to outline the need for derived data products, describe how the community can contribute to the determination and derivation of data products from NEON data, provide a few existing examples to show the process from beginning to end for creating derived data products, and address challenges and opportunities moving forward. This manuscript may serve as a guide to the creation of derived data products from any data source, public or private, but given the ubiquity and broad use of NEON data, we will specifically focus on NEON data.

The benefits and uniqueness of NEON data

Creating derived data products from NEON data requires a strong understanding of how NEON data are collected, maintained, and organized. NEON is a centrally managed, top-down network, enabling a level of standardization not often possible in more grassroots scientific measurement networks (Hinckley et al., 2016; Kao et al., 2012). Standardization of data collection and data products across both space and time ensures data integrity and interoperability that enables large-scale analyses and change detection. NEON sites were selected to facilitate analyses of cellular, organismal, and ecosystem processes across gradients occurring at all spatial scales, from site level to continental (Schimel et al., 2007). NEON divides the North American continent into 20 unique eco-regions, or "domains" (16 in the contiguous United States, and 4 more in Alaska, Hawaii, and Puerto Rico). Domains are defined based on topography, climate, and soil properties, and were identified using a rigorous and repeatable multivariate statistical analysis (Hargrove & Hoffman, 1999, 2004). Each domain contains an average of four, and as many as six, sites, for a total of 81 sites across the network (47 terrestrial and 34 aquatic).

Data collection procedures are standardized across sites, facilitating comparisons across broad spatial scales, and enabling robust time series analyses. Within each site, co-located climate, biodiversity, biogeochemical, and hydrological data are collected, enabling analyses and models to incorporate timely data about both driver and response variables, and to integrate data from diverse components of the ecosystem. Data management and processing are also standardized and centralized at NEON, ensuring interoperability among datasets for use in models and algorithms, and optimizing data for automated pipelines (Nagy et al., 2021; Ordway et al., 2021).

The challenges of NEON data

The large number and diversity of NEON data serve to address a broad range of research questions yet create challenges in discoverability and usability. Researchers must first know which data products are available and relevant to the work they are doing, and then to use that data successfully, they must understand the data structures, metadata, and documentation for those data. The NEON Data Portal (https://data.neonscience.org/) is the primary venue for access to NEON data. Through the Data Portal, users may query by site, state, ecological domain, data theme, or timeframe. For the uninitiated, learning to work with the NEON Data Portal may be likened to the challenges that previous generations faced trying to find books in a vast research library, where the books are organized using an unfamiliar classification system, and the card catalog is handwritten in an unfamiliar script.

NEON data are grouped by science themes: Atmosphere, Biogeochemistry, Ecohydrology, Land Cover & Process, and Organisms, Populations, and Communities. Within these themes, there are three major types of NEON data: (1) instrumental, (2) observational, and (3) remotely sensed. Instrumental and observational data are further divided into aquatic and terrestrial components. Structurally, these data vary widely, and often a given research community may be focused on or familiar with only one or two types of these data which can limit awareness and knowledge transfer. In addition, methods of data processing, formatting, and QA/QC procedures differ substantially among data types (Sturtevant et al., 2022), requiring data users to familiarize themselves with a broad array of data and data structures to work with the full range of available data.

NEON data are necessarily complex given the broad array of variables that must be curated, documented, and distributed using standardized methods to ensure data integrity and interoperability. The resulting data products vary in file size, format, and complexity and are designed primarily to be machine readable. Even relatively small NEON datasets are often far more complex and larger than datasets many ecologists are familiar with. This challenge is not necessarily unique to NEON data, highlighting the need for strong data science and programming skills across science and ecology (Borghi et al., 2018;

Nagy et al., 2021). To counter the perception that working with NEON data requires extensive coding experience (which may deter some researchers), NEON lowers barriers to entry by providing code packages for basic data access and wrangling (Lunch et al., 2020) and a Code Hub for members of the community to share their code (https://www.neonscience.org/resources/code-hub), as well as frequent workshops and courses (https://www.neonscience.org/resources/learning-hub/workshops-courses). However, even with these resources available, using NEON data can be challenging and can require skills development in the areas of data science and coding.

NEON data effectively represent a well-organized, but incomplete and imperfect set of observations to characterize ecosystems, as even though data are co-located and standardized, often measured at different scales or resolutions. For example, soil microbial sequencing, soil chemistry, and root traits are measured at the soil core level, while vegetation traits are measured at the individual level. As such, it is often necessary for users to wrangle, process, and synthesize these data for their own needs, which frequently leads to redundancy, with different research teams each performing their own calculations to create comparable products. However, we may choose to view this challenge as an opportunity for the research community to share datasets that originate with NEON data, thus lowering barriers to reproducible, innovative, large-scale research. Such collective efforts could enhance the usability of the data, both by providing numbers that are easier to scale up and to compare across space and time, and by providing simplified datasets that are easier to understand and use. It would also represent a new, innovative, and decidedly 21st-century approach to ecology.

The NEON data hierarchy

NEON data are organized into several hierarchical "levels" according to the stage of QA/QC and data type (Figure 2; Appendix S1: Table S1) following the data level heuristic established by NASA (https://www.earthdata.nasa.gov/engage/open-data-services-and-software/data-information-policy/data-levels). Level 0 data are unprocessed raw data and are typically not published. Level 1 data are quality-controlled and quality-checked (QAQC) raw measurements converted to relevant scientific units (e.g., barometric pressure, surface reflectance). Levels 2 and 3 data are temporally and/or spatially resolved (e.g., rate of change of CO_2 concentration, mosaicked rasters of remote sensing derived vegetation indices). Level 4 data are calculated from lower level data or involve multiple input data products (e.g., such as the net flux of carbon between the ecosystem and atmosphere [DP4.00067.001]; continuous measurements of stream height and intermittent estimates of discharge [DP4.00130.001]). Due to multiple factors, including the scope of NEON's mission and budgetary limitations, NEON publishes only a limited number of Level 4 derived data products, compared with what could be generated.

DERIVED DATA PRODUCTS

Defining derived data products

Derived data products are unique, curated, quality-controlled, documented, and reproducibly generated datasets that result from applying distinct and diverse processing steps such as cleaning, filtering,

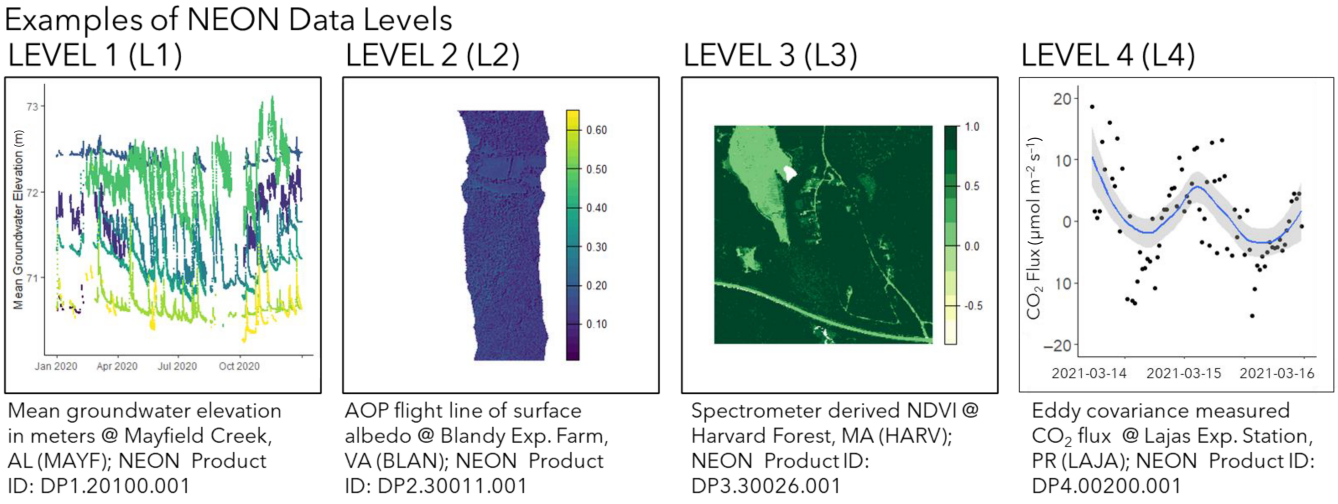

FIGURE 2 Illustrated examples of various data levels of National Ecological Observatory Network (NEON) data products, representing several NEON data themes.

transformation, fusion, spatial–temporal interpolation/extrapolation/integration, and/or statistical modeling to one or more primary data sources (Hofierka & Cebecauer, 2007; Peterson, 2005; Wan et al., 2017). Following the L1–L4 heuristic, derived data products are mostly like Level 4 or higher, as data become more highly derived as they move from Levels 0 to 4 (Table 1). The primary data for derived data products may include raw observational and measurement data at Level 0, L1–L4 data, or external, non-NEON data. Derived products involve the integration of multiple data sources, levels, and products—surmounting the limitations of existing NEON data products and providing insight and utility beyond those existing current products. Derived data products are also hosted in repositories other than NEON (e.g., EDI, ESS-DIVE, FigShare, Dryad, ORNL DAAC) given they are neither produced nor maintained by NEON.

Derived data products aim to extract valuable information, enhance data quality and accessibility, and provide customized data sets that address specific requirements, serving as tools to facilitate research rather than constituting a study themselves (Dwyer & Mason, 2018; Stanley et al., 2023). Derived data products can vary in scope—from local to nearly global in scale—and in structure—from tabular data files to highly structured data packages. For example, FLUXNET2015 and FluxnetEO offer platforms to study land–atmosphere flux at large spatial scales, including over 200 variables from regional flux networks across the globe including daily and annual integrals derived from 30-min flux data (Pastorello et al., 2020; Walther et al., 2022). Similarly, the PhenoCam network (https://phenocam.nau.edu/) enables investigations into the effects of climate change on vegetation phenology. PhenoCam includes data from over 700 sites and an archive of 60 million images, and data products characterizing vegetation color on a daily time step, as well as seasonal transition date derived from those data (Moon et al., 2022; Richardson et al., 2018). However, derived data products need not be global or continental in scale to be vital. The LAGOS-NE dataset includes 17 northeastern and midwestern US States and 5 tribal areas, fostering research on water quality in freshwater systems at multiple spatial and temporal scales (Soranno et al., 2017). Derived products may also be specific to unique experiments such as the *fortedata* R package, which provides detailed data from an ongoing experimental forest disturbance manipulation in northern Michigan, serving as a resource for examining the influences of disturbance on forest carbon cycling (Atkins et al., 2021) or describe single sites over time such as ecological change from land use patterns at Coweeta Hydrologic Laboratory in North Carolina (Wurzburger et al., 2023) or forest productivity in response to chronic acidification at Fernow Experimental Forest in West Virginia (Adams et al., 2020).

There is already existing precedent and strong community effort behind the creation of derived data products from NEON data. For instance, the NEON Tree Crowns data set includes information on height and crown area of more than 100 million individual trees in 37 NEON sites (Weinstein et al., 2021). Further, NEON image data have been integrated into the PhenoCam Network and flux data from NEON towers are available via the AmeriFlux network (https://ameriflux.lbl.gov/). Additionally, NEON-DICEE offers stable isotope ratios of net fluxes from tower profiles of atmospheric water vapor and CO_2 at various NEON sites (Finkenbiner et al., 2022). The continued development of derived data sets is essential to increase the size and diversity of the NEON user base and stimulate cross-disciplinary research initiatives.

TABLE 1 NEON data levels (adapted from https://www.neonscience.org/data-samples/data-management/data-processing).

Data level	Description
Level 0 (L0)	Raw sensor readings or human-made observations obtained in the field, for example, the 1 Hz resistance reading of a platinum resistance thermometer, or the species identification of individual plants along a transect.
Level 1 (L1)	Raw measurements are quality controlled and converted to relevant scientific units (e.g., Ohms, degrees Celsius). Measurements are often averaged to longer temporal or spatial scales and accompanied with aggregation statistics.
Level 2 (L2)	Temporally interpolated measurements or AOP data provided by flightline.
Level 3 (L3)	Spatially interpolated or mosaicked measurements, for example, 1 km tiles of vegetation spectral indices from AOP data.
Level 4 (L4)	The combination of basic measurements and scientific theory to derive higher order quantities. Examples include the computation of stream discharge from surface water elevation and a stage-discharge rating curve, and the exchange of carbon dioxide between the surface and the atmosphere from high-frequency wind and gas concentration measurements.

Abbreviation: AOP, Airborne Observation Platform.

The necessity of derived data products

In the era of big data, the generation and collection of vast datasets have become an integral part of scientific

research. NEON is a prominent example of a continental-scale ecological observatory that has amassed a wealth of raw data across diverse ecosystems (Nagy et al., 2021). Large datasets, in their raw form, can be challenging to work with for many end users. It may take valuable time and resources to transform those data into comprehensible, useful formats (Lohr, 2014; Wickham et al., 2019), thus researchers may expend considerable and duplicate effort cleaning and transforming data (O'Brien et al., 2021) or represent processes in noncomparable ways. This redundancy can be mitigated through the provision of quality-controlled derived data products that allow scientists to focus their efforts on hypothesis testing and analysis rather than data preparation—saving not only time and resources, but also promoting accuracy and reproducibility (Li et al., 2022). By providing high-quality, standardized datasets, the NEON user community can enable researchers to build upon each other's work, facilitating open and reproducible science (Balch et al., 2020). These derived products contribute to the cumulative nature of scientific knowledge, making it easier to replicate experiments, validate findings, and draw robust conclusions. Therefore, the availability of processed data accelerates the pace of research, enabling scientists to respond more quickly to emerging environmental challenges. Additionally, efforts to produce well-documented data products that comply with FAIR principles (see *FAIR and open data principles*) boost the reputation of the data creator(s), provide quantitative benefits in the form of citations, and use metrics that are part of performance and tenure review (Colavizza et al., 2020), support data sharing expectations from funding agencies given their highly complementary with what is expected in most open science/data management plans (White House Office of Science and Technology Policy [OSTP], 2022), and elevate future funding proposals if incorporated explicitly.

Derived data products may serve as links between disparate datasets, enabling researchers to identify and explore related data with common themes. For instance, categorizing and indexing derived products can facilitate the discovery of adjacent datasets that may not be immediately apparent, thus encouraging interdisciplinary collaborations among researchers from various disciplines (Tobi & Kampen, 2018). The availability of processed data further facilitates education and training, allowing educators to incorporate real-world, up-to-date environmental data into their classroom—promoting environmental literacy and inspiring the next generation of scientists and policy makers.

FAIR and open data principles

For derived data products to facilitate interoperability and discovery-based science, alignment with good data management practices during production is a necessity. Four foundational principles (Findability, Accessibility, Interoperability, and Reusability)—the FAIR data principles—have become the standard practice for increasing the reusability of data products (Wilkinson et al., 2016). NEON derived data products can be Findable and Accessible when published with a persistent DOI in an open repository (Lin et al., 2020) with reference, via citation and inclusion of the DOI for the input NEON data product(s). Derived data products registered with DOIs can (and should) be cited in subsequent manuscripts, proposals, and reports using those data, contributing to citation counts. However, data citation practices differ across journals and disciplines (Robinson-García et al., 2016; Silvello, 2018) currently creating often incomplete portraits of the contributions of those data products—although attitudes and practices are shifting. Derived data products can be Interoperable and Reusable by following community standards for reporting and formatting, including following community-defined metadata standards (Crystal-Ornelas et al., 2022; Poisot et al., 2019). However, maturity of community standards may vary dramatically by discipline or subdiscipline, and development of coordinated derived products may be a reasonable way to increase standardization of datasets by involving researchers and practitioners with diverse skill sets in their production (Poisot et al., 2019). The interoperability of derived data products can be maximized when published with links to the underlying input datasets via appropriate citation (including persistent identifiers, e.g., DOIs and direct citation of the NEON data used in that product as well), their production scripts, and metadata.

Well-curated derived data products are often time-consuming and difficult to produce, particularly given the broad data types provided by large research and observatory networks such as NEON. Production of derived data products from NEON datasets by and for the broader community therefore benefits from using holistic approaches throughout the research life cycle such as the ICON (Integrated, Coordinated, Open, and Networked) approach (Goldman et al., 2022). The ICON approach embraces open coordination and collaboration by design, where considerations and collaboration around data development and standardizations occur continuously from data collection to publication (Goldman et al., 2022). Derived data products may speed the rate of scientific advancement by creating greater capacity for knowledge synthesis across key data types and disciplines

(Dwivedi et al., 2022). For example, reproducible approaches that integrate spatially and temporally mismatched datasets across and within NEON sites can enhance our ability to address research questions across spatial and temporal scales (Meier et al., 2023). Greater mutual benefit across scientists, stakeholders, and elsewhere can be facilitated through networked production of derived data products (Jones and Nelson, 2021): for example, production of derived data products spanning coordinated phenological datasets from NEON and the USA National Phenology Network's Nature's Notebook (Denny et al., 2014; Elmendorf et al., 2016) may enhance mutual benefit for involved stakeholders (Dwivedi et al., 2022).

Data licensing

Creators of derived data products should strongly consider releasing their data products under a license. Licenses protect both the creators and users of the data products from copyright issues that may arise from the use, distribution, reproduction, or modification of the derived data products. Moreover, default copyright laws may inhibit data use and distribution more than a researcher intends. For example, under US copyright law, all rights are reserved to the creator of the work by default. Because of this, default assignment of rights, derivative works, or works with multiple creators may unwittingly result in unexpected limitations on data use or copyright violations (for more info on licensing see the Center for Open Science: https://help.osf.io/article/148-licensing; or the How to FAIR website: https://howtofair.dk/how-to-fair/data-licences/).

The most used licenses for data are the Creative Commons (CC) Licenses, with several versions that place different conditions on data use, derivative works, and attribution. The most permissive license is CC0, which places products in the worldwide public domain. NEON releases its data products through the CC0 license. Beyond CC0, CC maintains six additional licenses that place additional restrictions on data use: attribution (BY), share alike (SA), no derivative works (ND), and no commercial use (NC). CC BY allows distribution, derivatization, and reuse so long as credit is given to the creators of the original dataset. CC BY-SA and CC BY-NC are modified versions of CC BY that require derivative works to use the same license and prohibit commercial use, respectively. CC BY-NC-SA applies to both restrictions. Finally, two additional license versions prohibit derivative works: CC BY-ND allows reuse of the data for any purpose with attribution, indications the data were changed, and a restriction on distribution of any modified version of the data, while CC BY-NC-ND allows only noncommercial reuse. We recommend that researchers seeking to maximize the community use of their data consider one of the more permissive licenses (e.g., CC BY-SA, CC-BY, or CC0; creativecommons.org).

DEVELOPMENT, DOCUMENTATION, AND DISTRIBUTION

Developing derived data products

The first step in the development of a derived data product is to identify a specific need. Most commonly, ideas for derived data products emerge organically from ongoing research efforts and reflect the needs of the researcher or research group producing them. However, ideas for derived data products can also be in response to a broader identified community need. Regardless of impetus, it is important to consider community needs and standards from the outset to ensure that the final derived data product is useful to as many end users as possible (Boxes 1 and 2).

An effective way to ensure that a derived data product is broadly useful is by incorporating a diverse array of team members. It may be prudent to reach out to other researchers working with the same raw data to identify any potential synergies and to avoid duplicate efforts. Researchers should not hesitate to reach out to NEON data scientists who are available to help with technical assistance and information on how NEON data are used and possible community needs. In addition, derived data products can vary in their difficulty of production. Some derived data products can feasibly be produced by a single researcher in a few days or weeks while others may require cross-disciplinary efforts and various domain expertise and data science skills and take months or longer to develop. In building collaborations, it is important to consider looking beyond your immediate network to intentionally include historically underrepresented groups and researchers at different career stages. When larger teams are required, best practices for team science (Cheruvelil & Soranno, 2018) can help guide the collaboration. Additionally, the aforementioned ICON approach can also be consulted (Goldman et al., 2022). At a minimum, it is important to set a realistic timeline and assign specific roles and responsibilities to team members. A conventional approach for this could include a collaborative governing document, which outlines individual roles and responsibilities for the duration of the project as well as specific procedures for addressing conflict and authorship within the working group

BOX 1 Useful questions/considerations regarding the creation of derived data products.

1. How will FAIR principles be addressed?
2. What is the plan for finding resources to sustain the product after the data product is completed and who will handle sustainment?
3. How will the product initially be reviewed for quality and completeness?
4. Are partnerships necessary to complete the minimum viable product?
5. What existing community network of experts will be consulted regarding the design and distribution of the product?
6. What repository will be used and what are its requirements including data and metadata, size, funding source, versioning, and topic of the product?
7. If you are a junior researcher, who is the mentor or adviser who will provide direction and review of the plan?
8. What level of completion or quality control is minimally necessary to provide an initial demonstration of the product's usefulness, and how much effort will it take to get there?
9. How can the accessibility, usefulness, and adoption of the product by the widest possible audience, including users outside your primary audience, be maximized?
10. What sort of training or documentation is needed to make the product useful?

BOX 2 An approximate order of operations for consideration in the timeline of creating a derived data product.

1. Identify the need

First you must know what data product is needed. This can be based on an obvious gap in what is available in your specific research community or a common need shared among researchers.

2. Build the team

Consider the necessary technical and domain knowledge needed and make sure to consider representation explicitly. Diverse teams create stronger, better results. Use Team Science Resources to help guide the process. Remember NEON data scientists and staff as resources.

3. Planning

Consider the necessary technical and domain knowledge needed and make sure to consider representation explicitly. Diverse teams create stronger, better results. Use Team Science Resources to help guide the process. Remember NEON data scientists and staff as resources.

4. Science

Begin! Consider starting with a subset of your data and create a "prototype" data product. This will help in identifying issues earlier and create a more efficient process. Then proceed. Make sure to comment and document. Your data are only as good as your process and your process is only as good as your documentation.

5. QA/QC

Check, test, verify, and repeat. The QA/QC process may be the most time-consuming but is just as important as the documentation and planning. Specifics may vary among data or disciplines, but generally include checks for data consistency, outlier detection, verification, etc.

6. Distribution

After your data product passes QA/QC checks, it is time to upload it to a repository. Each repository has metadata requirements to document the data product and do vary slightly. Adhere to all of them and be specific. Now it is time to write that data paper and tell all your friends!

(Baumgartner et al., 2023). For more structure, a formal working group could be formed (e.g., through the NSF-funded Environmental Data Science Innovation and Inclusion Lab at University of Colorado https://esiil.org/working-groups).

After building a diverse team with the required expertise, it is time to plan the workflow of the project. When creating derived data products, it is necessary to be well-organized and intentional. The workflow should be as well documented and planned out as the final product. Each step in the process from raw data to finished product requires an intentional choice, whether it be specific algorithms, conversion factors, coding platforms, responsible personnel, or data repositories to house the data.

Begin with a smaller scale prototype or case study (e.g., using data from just one NEON site or collection year) to establish an analytical workflow (Stoudt et al., 2021). This allows for the development of the overall framework, but without computational constraints that may hinder full-scale implementation. For example, this step may involve identifying the necessary input data sets, the types of QC and cleaning that these inputs require, the best processing methodologies, and an appropriate QC and validation step of the end product. Once development of a smaller scale prototype is completed, the computational requirements and time commitment for the full analysis can be better assessed. Finally, the complete derived data product can be produced, likely following many of the steps detailed in Figure 3 and Box 2.

Documenting derived data products

Documentation is a central consideration. A key advantage of using NEON data is the accompanying documentation, typically in the form of an Algorithm Technical Basis Document (ATBD) or sampling protocol. ATBDs are outstanding examples of documentation and should be consulted during the process of using NEON data or in creating derived data. Documentation should also extend to all code and scripts used to generate a data product. Code documentation should be written following community guidelines and include comments, consistent formatting, and be made fully open and accessible on platforms like GitHub. Derived data products should be accompanied by user-friendly documentation such as vignettes or tutorials—written using Markdown, Jupyter Notebooks, Bookdown, or similar platform. Jupyter notebooks and JupyterLab (jupyter.org) provide web-based interactive environments for running and documenting code and analyses, with the capability of integrating multiple coding environments (e.g., Julia, R, Python). R Markdown offers a straightforward and powerful means of documenting data products in the R software and statistical environment, a common tool in ecology-related disciplines (Atkins et al., 2022). These

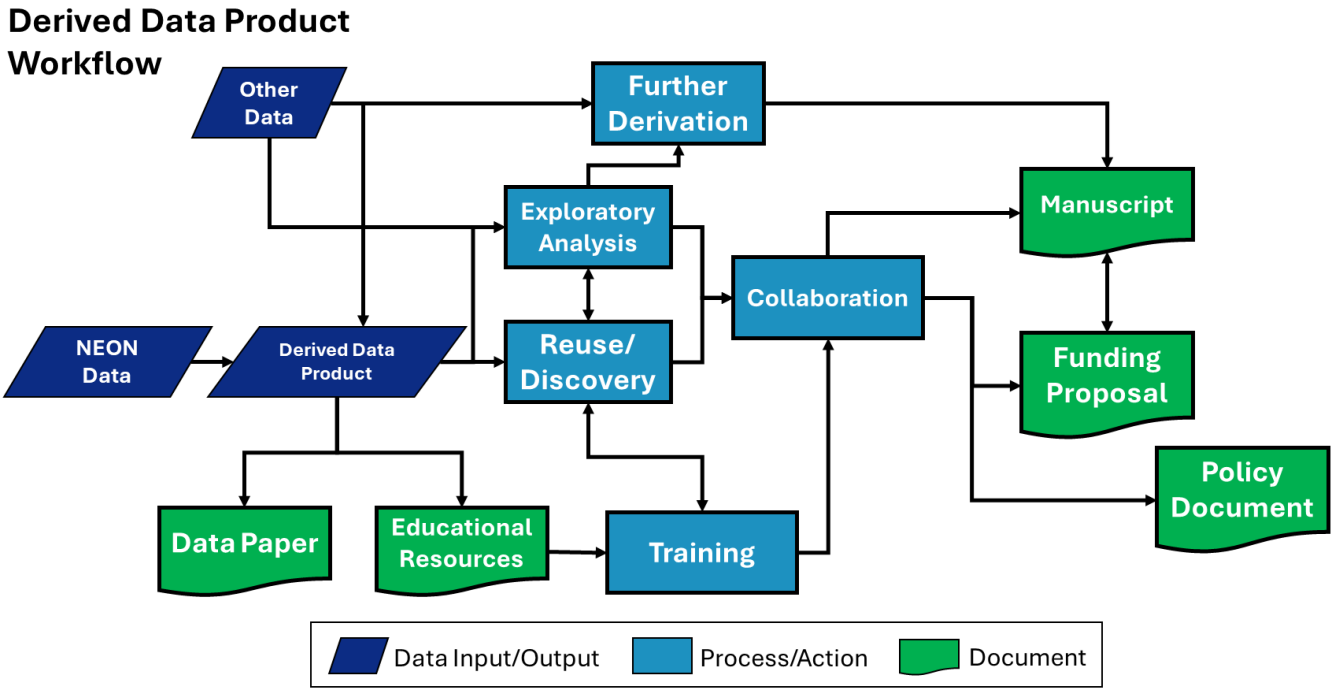

FIGURE 3 Flowchart depicting the interconnected nature of derived data products.

platforms seamlessly blend executable code, outputs, and descriptive text into cohesive, human-readable documents. They present code in a format enriched with example outputs like figures and maps, allowing researchers to visualize and comprehend the process and outcomes concurrently. The interactive nature of R Markdown documents and Jupyter notebooks, featuring dynamic code chunks and their real-time results, enhances user engagement, promoting a hands-on approach to learning and exploration. README files are a particularly necessary form of documentation to accompany code and should include instructions on how to work with the included code and detail installation if organized in a package structure or similar. Workflows are necessary to allow an immediate overview of how the code works and what it does to make it easier for researchers to reproduce or build upon the existing work. The README fosters an environment of transparency and inclusivity, enabling diverse users, irrespective of their familiarity with the project, to engage with, adapt, and build upon the existing work. The individual platform used to create documentation may vary by coding environment, product complexity, or creator familiarity, but resulting documentation should be easy to view, read, understand, and share. Existing NEON education resources provide documentation examples: (https://www.neonscience.org/resources/code-hub).

Graphics can be essential in helping to describe a data product. A well-structured workflow diagram can offer readers a visual representation of the method. Additionally, a clear outline of the format of all files and records, specifying units, variable names, and detailing any other relevant attributes ensures that users are well-equipped with the necessary information to effectively utilize the dataset in their work. Documenting the quality of the dataset is another important step. Detailed technical validation or QA/QC checks on derived products can reassure researchers of their reliability, accuracy, and credibility. This may be as informal as a description of filtering or outlier detection methods employed for data removal or as formal as testing procedures (e.g., "testthat" package in R).

Versioning derived data products

It is also important to provide versioning on both the derived data products and associated code (Crystal-Ornelas et al., 2022). Common versioning approaches involve maintaining a detailed changelog with each version, documenting any alterations, enhancements, or fixes, thereby ensuring transparency and reproducibility. Leveraging established version control systems like Git, in conjunction with platforms such as GitHub or GitLab, enables seamless tracking of changes and facilitates collaboration. Additionally, associating each version with a unique DOI ensures that every iteration is easily citable and accessible, promoting academic integrity and proper referencing (Crystal-Ornelas et al., 2021). For example, data repositories (e.g., ESS-DIVE, figshare) provide DOIs for each published data set or update to a data set for all data products, and individual DOIs for new code releases in a Github repository can be created using Zenodo.

Hosting and distributing derived data products

To distribute a derived data product once it has been created, it is necessary for it to be hosted somewhere where it is publicly available and downloadable. As previously outlined, derived data products are only likely to be adopted and used by the community when they are published following FAIR principles (Wilkinson et al., 2016). Rich metadata, persistent identifiers, and open source code alone will not guarantee intended audiences will find data products. Findability, therefore, requires the publication of data products in widely used and accessible repositories so that they are more discoverable. For example, the inclusion of NEON-specific keywords (e.g., "National Ecological Observatory Network", NEON data product IDs, NEON four-letter site codes) and DOIs from NEON data releases (https://commons.datacite.org/repositories/x3ockqg) improves findability. NEON provides a set of guidelines and best practices for publishing research products, including derived data products (https://www.neonscience.org/data-samples/guidelines-policies/publishing-research-outputs). NEON partners with the Environmental Data Initiative (EDI), which serves as a repository for a wide variety of environmental datasets and provides resources and support for submission of user-created datasets for publication. Datasets in the EDI repository are searchable based on spatial, temporal, and taxonomic factors and are accessible through the DataOne portal and Google Dataset Search. NEON plans to incorporate enhanced discoverability of NEON-related data sets shared via EDI on its website in the future. EDI supports data of many types and formats, including tabular, spatial, image/video, and netCDF, as well as code. EDI requires the use of the Ecological Metadata Language (EML), an xml metadata standard, for all metadata associated with the "data package." The elements that must be included in the EML files include information on the content (e.g., title, abstract, keywords), creators, and spatial and temporal coverage of the data, as well as the methodology by which the data were created. EDI provides an online form application

(ezEML) to facilitate creation and submission of the EML files and NEON provides a NEON-specific EML template (https://www.neonscience.org/sites/default/files/NEON_EML_Template.zip) that pre-populates where and how NEON should be referenced (e.g., listed as a data provider in the Associated Parties section, included in the abstract through an acknowledgment statement). Following submission, the data package is reviewed for validity of the metadata and published to the repository with a DOI, after which the author can make versioned updates and track downloads and citations of the data package.

While EDI may be suitable for most NEON derived data products, publication to additional or better-suited discipline-specific repositories may often be necessary. For example, derived data products that can be tied directly to samples or specimens collected by NEON (e.g., community metrics for bulk samples, trait data derived from specimen images) should be submitted to the NEON Biorepository data portal (https://biorepo.neonscience.org/), where observations will be associated with occurrence records following DarwinCore data standards (Wieczorek et al., 2012), pushed to the Global Biodiversity Information Facility (GBIF) data portal, and published as citable datasets on EDI. Some derived data products may be rather large and require coordination with data repositories or even cost-sharing agreements to facilitate hosting. Other options for large data products may include distribution through Distributed Active Archive Center or DAACs such as the Oak Ridge National Laboratory (ORNL DAAC) or via private solutions such as Google Earth Engine Catalog (Appendix S1).

Promoting and publicizing derived data products

Following documentation, creation, and publication, the next step is to promote and use the data product. Even the most useful and well-documented products may sit untouched unless potential user groups know that they exist and can envision their potential applications. It is also important that the products are both easy to find and easy for the target audience to use. The simplest method for promoting a derived data product, especially if it is a product of ongoing research efforts, is to cite both the data product itself and the underlying analytical workflow in research publications. In addition, publishing regular data-product or code updates/versions as new data become available from NEON will help keep the derived data product current. Researchers can also reach out to NEON if they think a NEON Observatory Blog post (https://www.neonscience.org/impact/observatory-blog) may be of interest to the wider NEON community and promote the product via social media platforms or forums.

Writing a data paper in support of data products

We believe that the publication of a data paper that more fully details the need, derivation, and use-cases of any derived data products is a necessity. Data papers serve to not only document a data product but also advertise the product's utility to the research community. Of course, creating a data paper is optional, but any data product of reasonable complexity would be greatly enhanced by a data paper. The data paper should include a worked example(s) that allows readers to understand the practical application and potential of the dataset(s). Lastly, include a summary or conclusion that outlines how the data fit into the larger scientific community and include additional potential use-cases or applications. Several journals such as *JGR-Biogeosciences* and *Ecology* offer specific data paper formats, while other journals such as *Methods* X, *Scientific Data*, and *Earth System Science Data* specifically focus on data and data products; the list continues to grow as data papers become more common practice.

Regardless of outlet, the creation of a data paper mirrors the creation of a derived data product, beginning with the need to clearly identify and articulate the unique knowledge gap or need that the data set addresses within the scientific community—further highlighting the relevance and significance of the derived data product. Following this, it is essential to provide a comprehensive description of the raw data sources, NEON or otherwise, and the processing steps (e.g., algorithms, equations) involved in the transformation of the original data into the derived product. Be sure to include and cite any digital object identifiers (DOIs) associated with the input data—whether from NEON or elsewhere—as this maintains the provenance of the data, helps with reproducibility, and credits the input data producers as well. While manuscripts written in support of derived data products increase the time and effort invested in the process, they often garner additional citations and represent accomplishments that traditionally count toward promotion and advancement—especially for researchers in academic positions—and boost career visibility.

Education and classroom activities can also be good venues to encourage the reuse of the data product. For example, programs like Project EDDIE (https://serc.carleton.edu/eddie/index.html) and Data Nuggets (https://datanuggets.org/) provide guidance for creating teaching

materials around environmental data sets and venues to share these lesson plans with educators. In addition, creators of derived data products could reach out to instructors and educational institutions that have an environmental data science program (e.g., Earth Lab at the University of Colorado, the Ecological and Environmental Informatics PhD program at Northern Arizona University and similar programs at UC Davis, University of Virginia School of Data Science, Virginia Tech, Colby College, and Denison University, among others) to learn about what kinds of materials, exercises, or modules that could be useful for educational purposes in classes or training programs. Further, NEON has a wide array of education partners, and creators of derived data products are encouraged to reach out to them.

Maintaining derived data products

In addition to the short-term challenges faced in creating derived data products, creators should consider data product maintenance. Unlike scientific papers that represent a snapshot in time, there are expectations that datasets (and accompanying code) will be kept up-to-date and function in perpetuity—regardless of changes in style or software, although Docker containers (https://www.docker.com/) can be used to simplify development and deployment across platforms. Curation and maintenance are major challenges and considerations in the creation of derived data products, and it is necessary to plan for how data products will be kept up-to-date and how any community comments, revisions, suggestions, or improvements can be responded to and incorporated when appropriate. In the case of NEON data, new data are added at regular intervals, necessitating consistent update for derived data products to remain up-to-date and relevant. Occasionally, revisions to Level 0 data may occur, necessitating reprocessing of a derived data product. In either case, the researcher or group who created that product may not have the ability to reprocess the data product either through funding limitations or loss of skills due to personnel turnover—it should be noted that while it is infeasible to expect updates of every product after funding has ended, this does not necessarily mean those data products are useless!

EXAMPLE DATA PRODUCTS

The following section provides detailed examples of recent derived data products created from NEON source data with the objective of demonstrating the process behind the creation of data products (Figure 3, Box 2), including the demonstrated need, establishment of workflows, adherence to FAIR principles, and discussion of the challenges confronted and how they were surmounted.

Example 1: Forest biomass from inventory and allometry

Aboveground biomass is a recognized Global Climate Observing System (GCOS) Essential Climate Variable (ECV) with accurate and constrained biomass estimates important for understanding the global carbon system, informing carbon offset and biomass markets, monitoring aboveground carbon storage, and driving Earth system models (Herold et al., 2019). Aboveground biomass estimates for forests are predominantly made at the individual tree level using stem diameter measurements and allometric scaling functions—species, genus, and/or regionally specific equations developed via destructive sapling relating stem diameter measurements with overall tree volume and biomass. Biomass estimation via allometric scaling can be straightforward but does require some level of domain and statistical knowledge—issues compounded by the sheer number of trees in the NEON database.

To assess this need, Atkins et al. (2024) created a derived data product, the *NEONForestAGB* data set and R data package, which includes individual tree-level biomass estimates for 91,390 live stems located in a total of 1233 observation plots at 40 NEON terrestrial sites from 2016 to 2022. Biomass estimates were made from NEON vegetation structure data (DP1.10098.001; NEON, 2023) and processed in R (R Core Team, 2023) using the *neonstore* package (Boettiger et al., 2021), with custom functions and lookup tables created by the team. Lookup tables included allometric scaling coefficients from two separate allometries from external data sources (Chojnacky et al., 2013; Jenkins et al., 2003). Each of these allometries is a generalized allometry, meaning it is specific at the family or genus level and not regionally specific. For example, Jenkins et al. (2003) included 10 model forms (e.g., "pine," "fir/hemlock," "soft maple/birch") while Chojancky et al. (2014) included 34 model forms, based largely on the Jenkins et al. (2003) groupings but providing further discretization based on specific wood gravity. Atkins et al. (in review) estimated biomass using a logarithmic function with an additional step taken to adjust estimates for saplings to include a taper function and biomass inflation adjustment factor. The subsequent derived data product then includes all the necessary identifying information for each tree (e.g., site, plot, species, diameter) and two estimates of biomass

based on each allometry. The creation of the *NEONForestAGB* product involved collaboration among domain experts, such as botanists and forest ecologists, as well as data science and coding experts, including NEON staff scientists who were instrumental in the process. *NEONForestAGB* addresses the need of the community to have standardized biomass estimates from NEON forest inventory data that can be used for research and management purposes. The data product has already benefited the team and their collaborators, supporting manuscripts and grant proposals.

Example 2: Greenhouse gas concentrations in inland waters

Inland waters (e.g., streams, rivers, and lakes) play important roles in carbon and nitrogen cycles due to their high rates of biogeochemical activity. Some of these biogeochemical transformations produce greenhouse gases (e.g., carbon dioxide, methane, and nitrous oxide), making inland waters natural sources of these climate-relevant gasses. However, there are large uncertainties in the estimates of greenhouse gas production and emission from inland waters, in part due to a limited number of direct measurements of dissolved gas concentrations.

To measure greenhouse gases in inland waters, NEON collects "headspace equilibration" grab samples at their aquatic sites at a temporal scale of either ~12 (lakes) or ~26 (streams and rivers) times per year. The headspace equilibration method used by NEON involves equilibrating a water sample and an air headspace in a syringe and then analyzing the equilibrated headspace and a paired air sample on a gas chromatograph. NEON publishes raw mixing ratios for both the headspace sample and air sample on their Data Portal as the data product, Dissolved gases in surface water DP1.20097.001. However, these two raw mixing ratios are ecologically meaningless without significant data processing to convert them to a derived data product of either dissolved gas concentration or partial pressure.

A derived data product, published on EDI, Dissolved greenhouse gas concentrations derived from the NEON dissolved gases in surface water data product (DP1.20097.001), presents molar concentration and partial pressure of dissolved carbon dioxide, methane, and nitrous oxide for the 34 NEON aquatic sites through September 2020 (Aho et al., 2021, 2023). The workflow to create this product involved QCing the input files (e.g., identifying leaked and mislabeled vials, assessing reproducibility of field triplicates, logging all QC decisions in an issue log); identifying the best sources of required ancillary data (e.g., paired water temperature, barometric pressure, alkalinity concentrations); determining how to deal with missing values, flagged values, and values below the detection limit; writing and executing the processing scripts; and QCing the outputs of these scripts. The data set reflects the collaborative work of three separate research groups who were all independently using the NEON dissolved gas data set, all of which have since used the derived data product (Aho et al., 2023; DelVecchia et al., 2023; Stanley et al., 2023). Consultation with NEON data technical experts was important/critical in developing these derived data products. Updates will be required to keep this derived data product current as NEON continues to release raw mixing values.

Example 3: Remotely sensed road networks for NEON sites

Socio-ecological systems research explores linkages and feedbacks between human and natural components of ecological systems. This perspective is critical to understanding spatial patterns and dynamics over time at NEON sites. Early on, NEON was intended to have a fourth set of data products documenting land use and management across the network (i.e., the Land Use Analysis Package [LUAP]) to complement the Terrestrial Observation System, Aquatic Observation System, and Airborne Observation Platform (AOP) data streams (Keller et al., 2008). However, the LUAP data product was removed from NEON's scope to address budget issues during the Observatory's construction. NEON provides categorical land cover information provided by the National Land Cover Database (NLCD; Dewitz & USGS, 2021) derived from 30-m resolution Landsat imagery (Jin et al., 2023). NEON does provide site management and history data (NEON management and event reporting [DP1.10111.001]), yet knowledge of human system activities within landscapes surveyed by NEON is often lacking, particularly at the fine spatial scales afforded by the NEON airborne observation platform (AOP) remote sensing data.

One challenge of the coarse spatial resolution data provided by NLCD for NEON sites is that it excludes linear disturbances such as roads, paths, and railways, which are an increasingly common feature of many habitats. Linear disturbances can have important impacts on biodiversity through habitat fragmentation and the creation of edge effects (Forman & Alexander, 1998). Although the United States Census Bureau provides information on roads across the United States, their data layers focus on paved roads, and the exclusion of dirt

roads can lead to underestimation of the degree to which linear disturbances are present within a study area. To improve the quantification of linear disturbances at NEON sites, Record et al. (unpublished manuscript) used light detection and ranging (LiDAR) data from the NEON AOP to generate high spatial resolution (1-m) elevation maps of all NEON terrestrial sites. These maps were combined with road, railroad, and stream spatial layers from the United States Census and National Hydrography datasets (U.S. Geological Survey, 2019), respectively, and then shapefiles of additional, undetected linear features visible in the LiDAR map were made in QGIS. Final cumulative road and railroad layers for each NEON site were validated with high-resolution Google Earth imagery during leaf-off period in areas with deciduous canopy and additional archives of site-specific maps. The creation of this data product was an excellent opportunity for teaching undergraduates about team science, environmental data science, and the reproducible FAIR principle (e.g., use of GitHub to version and share R scripts for generating hillshade maps from NEON AOP LiDAR, documentation of versions of data downloads). Indeed, six undergraduate students and one post-baccalaureate research fellow contributed to this derived data product. The linear feature data product can subsequently be used for analyses of habitat fragmentation and edge effects at NEON sites.

Example 4: Community ecology data across networks

We are at an exciting time in ecology where there are multiple networks collecting ecological data across continents (e.g., NEON, the Long Term Ecological Research [LTER] Network [Jones & Nelson, 2021]; Australia's Terrestrial Ecosystem Research Network [TERN] [Cleverly et al., 2019], the South African Environmental Observation Network [SAEON] [Slingsby et al., 2023]; and the International Carbon Observatory System [ICOS], https://www.icos-cp.eu). Synergies across this network of ecological observatory networks have the potential to increase scientific insights made from their data streams (Jones & Nelson, 2021; SanClements et al., 2022). However, the challenges faced in working with data (e.g., interoperability) are multiplied when working across sites and research networks (Record et al., 2021). For example, community ecology data are inherently challenging to harmonize and make interoperable across data sets because of changing taxonomies across time and the sensitivity of inferences from the data to sampling techniques (O'Brien et al., 2021; Welti et al., 2021). As part of an LTER-EDI working group, O'Brien et al. (2021) created a data design pattern for such community ecology data, *ecocomDP* (O'Brien et al., 2021), with the intention of generating derived data products from LTER data sets to enable synthesis research. This led to further progress and scientific output, as during the 2019 NEON Science Summit held at the University of Colorado Boulder (Nagy et al., 2021), the developers of *ecocomDP*, and NEON staff experts associated with the various aquatic and terrestrial observation systems' data collaborated to format the NEON organismal data into the *ecocomDP* format. The effort took approximately two years with lots of remote communication between the team members (e.g., via GitHub issues). Generous time from NEON staff enabled the team to ensure that decisions made during the harmonization of the data appropriately documented the different organismal sampling methodologies in a meaningful way for later community ecology analyses (e.g., ordinations, beta diversity calculations; Jarzyna et al., 2022). The derived data product, *neonDivData*, was published as an R data package (Li et al., 2022) that is updated annually when NEON provides a new release version of their data. To build a community of users of the *ecocomDP* harmonized derived data products, including *neonDivData*, the developers frequently contribute workshops at national conferences on how to use the *ecocomDP* and *neonDivData* R packages.

CHALLENGES AND LESSONS LEARNED

The process of creating derived data products unearths many challenges, but also provides valuable lessons. Here we outline challenges encountered and lessons learned as gleaned from the teams that created the example cases detailed above (see *Example data products*).

Challenges faced

1. Appropriate reuse of data relies on investment in FAIR principles by data providers and users, which necessitates educating (and in some cases, reeducating) the community on the appropriate protocols and theory.
2. Inequities may exist where larger research groups with more access to resources may be more likely not only to create, but also be successful in distributing and having others use derived data products (Record et al., unpublished manuscript).

3. Creating quality data products is an investment of time and resources and may require creativity to meet those needs.
4. Creating derived data products can take time and sometimes team members graduate, transfer positions, or move onto other projects as the product is being made. Hence, it is essential that there is someone on the research team to champion organization of the project to make sure that methodology is reproducible.
5. There is currently insufficient credit structure within academic research to reward the amount of time and effort it takes to make a derived data product fully reproducible (Record et al., unpublished manuscript) or to support the continuous updates and engagement necessary to sustain a product.
6. Regarding updates and engagement to sustain a derived data product, there are insufficient funding mechanisms to support such endeavors.
7. It is important for NEON as an NSF large facility to track the use of its published data, including derivative products as well as manuscripts/papers/projects—all of which should cite both the source of the derivative product and the original NEON data source. This link will be lost if derived data products are cited without attribution to NEON data.
8. Technical issues may arise in the creation of data products including data gaps, incongruent timelines of data collection, or missing data necessary for certain derivations. This may result in the need for external data that must be collected or acquired, or new methods to overcome limitations.

Lessons learned

1. Generating derived data products from NEON's open data had the unexpected benefit of providing an opportunity to make research progress resilient during the COVID-19 pandemic (Record et al., 2022). For instance, the LiDAR-derived roads data product generation began during the pandemic when undergraduate students intended to participate in summer field research were unable to collect field data.
2. Collaboration with NEON scientists is vital for understanding the nuances of the data (Li et al., 2022). For example, in making the *neonDivData* tick data product, field and laboratory counts did not always coincide, so it was helpful to talk with NEON staff to make a sound decision as to what counts to include in the derived data product (Li et al., 2022). Similarly, NEON staff were instrumental in the creation of the *NEONForestAGB* product.
3. Derived data products create opportunities for integrating novice learners and students into applied research. Multiple points along the creation workflow (Figure 3, Box 2) create such opportunities, ranging from developing familiarity with data skills or even domain content.
4. Interdisciplinary collaborations tend to naturally spring from derived data product projects given the diverse range of required skillsets.
5. Documentation is and should be at the core of the entire process as it not only helps to describe the data product and increase its utility to the community, but it also serves as a valuable guide to team members.
6. Creating derived data products can be time-consuming but has profound benefits in bolstering the reputation of researchers beyond simple citation counts by creating future collaboration opportunities—particularly for early-career researchers.

CONCLUSIONS

Radical increases in the abundance and availability of environmental data have helped to usher in a new age of ecology and the natural sciences where we are poised to answer heretofore unapproachable questions. Yet this data revolution creates its own challenges, whether it be the need for skills to work with those data, or the effort expended in finding the right data. Advancing community efforts to synthesize extant data, such as those provided by NEON, creates one path forward to address these and other challenges we face. Here we have argued for the importance and viability of derived data products created from NEON data, and we have provided a framework, resources, and guide toward the creation of these data products. NEON thus creates abundant opportunities for the creation of many derived data products by researchers from across many disciplines. Through adherence to FAIR principles, recognized collective effort, the creation of diverse teams with varied skillsets, and listening to community needs, we can embrace the data-rich age in which we live.

ACKNOWLEDGMENTS

The conception of this manuscript sprang from participation by many of the authors in the National Ecological Observatory Network (NEON) Ambassador program and arose during discussions on building community engagement in the utilization of NEON data and resources. The findings and conclusions of this publication are those of the authors and should not be construed to represent any official USDA or US Government determination or

policy. NEON is a program sponsored by the National Science Foundation (NSF) and operated under cooperative agreement by Battelle. This manuscript is based in part upon work supported by NSF through the NEON Program. The Pacific Northwest National Laboratory (PNNL) is operated for the US Department of Energy (DOE) by Battelle Memorial Institute under contract number DE-AC05-76RL01830. Los Alamos National Laboratory (LANL) is operated by Triad National Security, LLC, for the National Nuclear Security Administration of DOE under contract number 89233218CNA000001. Andrew D. Richardson acknowledges support from NSF award number 2105828. Allison N. Myers-Pigg acknowledges workshop attendance support from PNNL's Program Development and Management and support for manuscript development from the DOE Office of Science (SC) Biological and Environmental Research (BER) program, as part of the Environmental System Science (ESS) Program to the River Corridors Science Focus Area at PNNL. Sydne Record acknowledges support from NSF award nos. 2301322 and 2242803 and the USDA National Institute of Food and Agriculture, Hatch Project award number ME0-22425 through the Maine Agricultural and Forest Experiment Station; Rich Fiorella acknowledges support from the Early Career Research Program of the DOE SC BER, Earth and Environmental Systems Science (EES) Division Earth and Environmental Systems Modeling Program as well as the Laboratory Directed Research and Development program of LANL under project number 20210961PRD3; Andrew J. Elmore acknowledges support from the Landscape Exchange Network for Socio-Environmental Systems (LENS) Research Coordination Network (RCN) project, NSF award number 2054939; Tong Qiu acknowledges support from NASA Earth Science Applications: Ecological Conservation, award number 80NSSC23K1534; Jeff W. Atkins acknowledges support from DOE-Savannah River Operations Office through US Forest Service-Savannah River under Interagency Agreement 89303720SEM000037. Benjamin Ruddell acknowledges support from the AccelNet-Implementation: Global Ecosystem Research Infrastructure (GERI), NSF award number 2301655. Danica Lombardozzi acknowledges support from NSF award number 2039932 and USDA NIFA award number 2021-0455. The authors would like to acknowledge E.A. Agee, P. Mabee, C. Nagy, E. Skipper, and K. Thibault for their valuable guidance in conceiving, drafting, and revising this manuscript. The authors also collectively acknowledge the time, effort, and thoughtful feedback provided by both anonymous reviewers and Kristopher Johnson who contributed to the revision and publication of this manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

No data were collected for this study.

ORCID

Jeff W. Atkins https://orcid.org/0000-0002-2295-3131
Xuan Chen https://orcid.org/0000-0002-9499-0054
Andrew J. Elmore https://orcid.org/0000-0002-9697-9457
Claire Lunch https://orcid.org/0000-0001-8753-6593
Luis X. de Pablo https://orcid.org/0009-0001-7911-4507
Allison N. Myers-Pigg https://orcid.org/0000-0002-6905-6841
Sydne Record https://orcid.org/0000-0001-7293-2155
Tong Qiu https://orcid.org/0000-0003-4499-437X
Samuel Reed https://orcid.org/0000-0003-3508-2547
Kelsey Yule https://orcid.org/0000-0002-1447-849X
Andrew D. Richardson https://orcid.org/0000-0002-0148-6714

REFERENCES

Adams, M. B., C. Kelly, B. Simpson, and J. Juracko. 2020. "Growth and Productivity of a 45-Year-Old Norway Spruce Plantation on the Fernow Experimental Forest." Page NRS-RN-253. U.S. Department of Agriculture, Forest Service, Northern Research Station, Newtown Square, PA.

Aho, K. S., T. Maavara, K. M. Cawley, and P. A. Raymond. 2023. "Inland Waters Can Act as Nitrous Oxide Sinks: Observation and Modeling Reveal that Nitrous Oxide Undersaturation May Partially Offset Emissions." *Geophysical Research Letters* 50: e2023GL104987.

Aho, K., K. Cawley, A. DelVecchia, E. Stanley, and P. Raymond. 2021. "Dissolved Greenhouse Gas Concentrations Derived from the NEON Dissolved Gases in Surface Water Data Product (DP1.20097.001) Ver 1." Environmental Data Initiative. https://doi.org/10.6073/pasta/47d7cb6d374b6662cce98e42122169f8.

Atkins, J. W., E. Agee, A. Barry, K. M. Dahlin, K. Dorheim, M. S. Grigri, L. T. Haber, et al. 2021. "The *Fortedata* R Package: Open-Science Datasets From a Manipulative Experiment Testing Forest Resilience." *Earth System Science Data* 13: 943–952.

Atkins, J. W., A. E. L. Stovall, and C. Alberto Silva. 2022. "Open-Source Tools in R for Forestry and Forest Ecology." *Forest Ecology and Management* 503: 119813.

Balch, J. K., R. C. Nagy, and B. S. Halpern. 2020. "NEON Is Seeding the Next Revolution in Ecology." *Frontiers in Ecology and the Environment* 18: 3.

Barnett, D. T., P. B. Adler, B. R. Chemel, P. A. Duffy, B. J. Enquist, J. B. Grace, S. Harrison, et al. 2019. "The Plant Diversity Sampling Design for the National Ecological Observatory Network." *Ecosphere* 10: e02603.

Barnett, D. T., P. A. Duffy, D. S. Schimel, R. E. Krauss, K. M. Irvine, F. W. Davis, J. E. Gross, et al. 2019. "The Terrestrial Organism

and Biogeochemistry Spatial Sampling Design for the National Ecological Observatory Network." *Ecosphere* 10: e02540.

Baumgartner, H. A., N. Alessandroni, K. Byers-Heinlein, M. C. Frank, J. K. Hamlin, M. Soderstrom, J. G. Voelkel, R. Willer, F. Yuen, and N. A. Coles. 2023. "How to Build up Big Team Science: A Practical Guide for Large-Scale Collaborations." *Royal Society Open Science* 10: 230235.

Boettiger, C., Q. Thomas, C. Laney, C. Lunch, and N. Ross. 2021. "neonstore: NEON Data Store."

Borghi, J., S. Abrams, D. Lowenberg, S. Simms, and J. Chodacki. 2018. "Support your Data: A Research Data Management Guide for Researchers." *Research Ideas and Outcomes* 4: e26439.

Cheruvelil, K. S., and P. A. Soranno. 2018. "Data-Intensive Ecological Research Is Catalyzed by Open Science and Team Science." *Bioscience* 68: 813–822.

Chojnacky, D. C., L. S. Heath, and J. C. Jenkins. 2013. "Updated Generalized Biomass Equations for North American Tree Species." *Forestry* 87(1): 129–151. https://doi.org/10.1093/forestry/cpt053.

Cleverly, J., D. Eamus, W. Edwards, M. Grant, M. J. Grundy, A. Held, M. Karan, et al. 2019. "TERN, Australia's Land Observatory: Addressing the Global Challenge of Forecasting Ecosystem Responses to Climate Variability and Change." *Environmental Research Letters* 14: 095004.

Colavizza, G., I. Hrynaszkiewicz, I. Staden, K. Whitaker, and B. McGillivray. 2020. "The Citation Advantage of Linking Publications to Research Data." *PLoS One* 15: e0230416.

Crystal-Ornelas, R., C. Varadharajan, B. Bond-Lamberty, K. Boye, M. Burrus, S. Cholia, M. Crow, et al. 2021. "A Guide to Using GitHub for Developing and Versioning Data Standards and Reporting Formats." *Earth and Space Science* 8: e2021EA001797.

Crystal-Ornelas, R., C. Varadharajan, D. O'Ryan, K. Beilsmith, B. Bond-Lamberty, K. Boye, M. Burrus, et al. 2022. "Enabling FAIR Data in Earth and Environmental Science with Community-Centric (Meta)Data Reporting Formats." *Scientific Data* 9: 700.

DataCite Commons. (2023). https://commons.datacite.org/ror.org/04j43p132.

DelVecchia, A. G., S. Rhea, K. S. Aho, E. H. Stanley, E. R. Hotchkiss, A. Carter, and E. S. Bernhardt. 2023. "Variability and Drivers of CO_2, CH_4, and N_2O Concentrations in Streams across the United States." *Limnology and Oceanography* 68: 394–408.

Denny, E. G., K. L. Gerst, A. J. Miller-Rushing, G. L. Tierney, T. M. Crimmins, C. A. F. Enquist, P. Guertin, et al. 2014. "Standardized Phenology Monitoring Methods to Track Plant and Animal Activity for Science and Resource Management Applications." *International Journal of Biometeorology* 58: 591–601.

Dewitz, J. 2021. National Land Cover Database (NLCD) 2019 Products (ver. 3.0, February 2024) [Data set]. U.S. Geological Survey. https://doi.org/10.5066/P9KZCM54

Dwivedi, D., A. L. D. Santos, M. A. Barnard, T. M. Crimmins, A. Malhotra, K. A. Rod, K. S. Aho, et al. 2022. "Biogeosciences Perspectives on Integrated, Coordinated, Open, Networked (ICON) Science." *Earth and Space Science* 9: e2021EA002119.

Dwyer, J. M., and R. Mason. 2018. "Plant Community Responses to Thinning in Densely Regenerating *Acacia harpophylla* Forest." *Restoration Ecology* 26: 97–105.

Elmendorf, S. C., K. D. Jones, B. I. Cook, J. M. Diez, C. A. F. Enquist, R. A. Hufft, M. O. Jones, et al. 2016. "The Plant Phenology Monitoring Design for the National Ecological Observatory Network." *Ecosphere* 7: e01303.

Finkenbiner, C. E., B. Li, L. Spencer, Z. Butler, M. Haagsma, R. P. Fiorella, S. T. Allen, et al. 2022. "The NEON Daily Isotopic Composition of Environmental Exchanges Dataset." *Scientific Data* 9: 353.

Forman, R. T. T., and L. E. Alexander. 1998. "Roads and their Major Ecological Effects." *Annual Review of Ecology and Systematics* 29: 207–231.

Goldman, A. E., S. R. Emani, L. C. Pérez-Angel, J. A. Rodríguez-Ramos, and J. C. Stegen. 2022. "Integrated, Coordinated, Open, and Networked (ICON) Science to Advance the Geosciences: Introduction and Synthesis of a Special Collection of Commentary Articles." *Earth and Space Science* 9: e2021EA002099.

Hargrove, W. W., and F. M. Hoffman. 1999. "Using Multivariate Clustering to Characterize Ecoregion Borders." *Computing in Science & Engineering* 1: 18–25.

Hargrove, W. W., and F. M. Hoffman. 2004. "Potential of Multivariate Quantitative Methods for Delineation and Visualization of Ecoregions." *Environmental Management* 34: S39–S60.

Heffernan, J. B., P. A. Soranno, M. J. Angilletta, Jr., L. B. Buckley, D. S. Gruner, T. H. Keitt, J. R. Kellner, et al. 2014. "Macrosystems Ecology: Understanding Ecological Patterns and Processes at Continental Scales." *Frontiers in Ecology and the Environment* 12: 5–14.

Herold, M., S. Carter, V. Avitabile, A. B. Espejo, I. Jonckheere, R. Lucas, R. E. McRoberts, et al. 2019. "The Role and Need for Space-Based Forest Biomass-Related Measurements in Environmental Management and Policy." *Surveys in Geophysics* 40(4): 757–778. https://doi.org/10.1007/s10712-019-09510-6.

Hinckley, E.-L. S., G. B. Bonan, G. J. Bowen, B. P. Colman, P. A. Duffy, C. L. Goodale, B. Z. Houlton, et al. 2016. "The Soil and Plant Biogeochemistry Sampling Design for the National Ecological Observatory Network." *Ecosphere* 7: e01234.

Hofierka, J., and T. Cebecauer. 2007. "Spatial Interpolation of Elevation Data With Variable Density: A New Methodology to Derive Quality DEMs." *IEEE Geoscience and Remote Sensing Letters* 4: 117–121.

Jarzyna, M. A., K. E. A. Norman, J. M. LaMontagne, M. R. Helmus, D. Li, S. M. Parker, M. Perez Rocha, et al. 2022. "Community Stability Is Related to Animal Diversity Change." *Ecosphere* 13: e3970.

Jeff Atkins. 2024. atkinsjeff/NEONForestAGB: NEONForest AGBv1.0.2 (Version v1.0.2) [Computer software]. Zenodo https://doi.org/10.5281/ZENODO.12795497

Jenkins, J. C., D. C. Chojnacky, L. S. Heath, and R. A. Birdsey. 2003. "National-Scale Biomass Estimators for United States Tree Species." *Forest Science* 49: 1–35.

Jin, S., J. Dewitz, P. Danielson, B. Granneman, C. Costello, K. Smith, and Z. Zhu. 2023. "National Land Cover Database 2019: A New

Strategy for Creating Clean Leaf-on and Leaf-Off Landsat Composite Images." *Journal of Remote Sensing* 3: 0022.

Jones, J., and M. P. Nelson. 2021. "Long-Term Dynamics of the LTER Program: Evolving Definitions and Composition." In *The Challenges of Long Term Ecological Research: A Historical Analysis*, edited by R. B. Waide and S. E. Kingsland, 55–79. Cham, Switzerland: Springer International Publishing.

Kao, R. H., C. M. Gibson, R. E. Gallery, C. L. Meier, D. T. Barnett, K. M. Docherty, K. K. Blevins, et al. 2012. "NEON Terrestrial Field Observations: Designing Continental-Scale, Standardized Sampling." *Ecosphere* 3: art115.

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* 6: 282–84.

Li, D., S. Record, E. R. Sokol, M. E. Bitters, M. Y. Chen, Y. A. Chung, M. R. Helmus, et al. 2022. "Standardized NEON Organismal Data for Biodiversity Research." *Ecosphere* 13: e4141.

Lin, D., J. Crabtree, I. Dillo, R. R. Downs, R. Edmunds, D. Giaretta, M. De Giusti, et al. 2020. "The TRUST Principles for Digital Repositories." *Scientific Data* 7: 144.

Lohr, S. 2014. For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights—The New York Times. https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

Lunch, C., C. Laney, M. Jones, and D. Durden. 2020. "Open Tools for NEON Data: Lessons from Open Code Development by NEON Scientists and the NEON User Community." ESS Open Archive, https://doi.org/10.1002/essoar.10501966.1.

Meier, C. L., K. M. Thibault, and D. T. Barnett. 2023. "Spatial and Temporal Sampling Strategy Connecting NEON Terrestrial Observation System Protocols." *Ecosphere* 14: e4455.

Metzger, S., E. Ayres, D. Durden, C. Florian, R. Lee, C. Lunch, H. Luo, et al. 2019. "From NEON Field Sites to Data Portal: A Community Resource for Surface–Atmosphere Research Comes Online." *Bulletin of the American Meteorological Society* 100: 2305–25.

Moon, M., A. D. Richardson, T. Milliman, and M. A. Friedl. 2022. "A High Spatial Resolution Land Surface Phenology Dataset for AmeriFlux and NEON Sites." *Scientific Data* 9: 448.

Nagy, R. C., J. K. Balch, E. K. Bissell, M. E. Cattau, N. F. Glenn, B. S. Halpern, N. Ilangakoon, et al. 2021. "Harnessing the NEON Data Revolution to Advance Open Environmental Science With a Diverse and Data-Capable Community." *Ecosphere* 12: e03833. https://doi.org/10.1002/ecs2.3833.

NEON Dimensions. 2023. https://neon.dimensions.ai/discover/publication.

O'Brien, M., C. A. Smith, E. R. Sokol, C. Gries, N. Lany, S. Record, and M. C. N. Castorani. 2021. "ecocomDP: A Flexible Data Design Pattern for Ecological Community Survey Data." *Ecological Informatics* 64: 101374.

Ordway, E. M., A. J. Elmore, S. Kolstoe, J. E. Quinn, R. Swanwick, M. Cattau, D. Taillie, et al. 2021. "Leveraging the NEON Airborne Observation Platform for Socio-Environmental Systems Research." *Ecosphere* 12: e03640.

Parker, S. M., and R. M. Utz. 2022. "Temporal Design for Aquatic Organismal Sampling across the National Ecological Observatory Network." *Methods in Ecology and Evolution* 13: 1834–48.

Pastorello, G., C. Trotta, E. Canfora, H. Chu, D. Christianson, Y.-W. Cheah, C. Poindexter, et al. 2020. "The FLUXNET2015 Dataset and the ONEFlux Processing Pipeline for Eddy Covariance Data." *Scientific Data* 7: 225.

Peters, D. P., P. M. Groffman, K. J. Nadelhoffer, N. B. Grimm, S. L. Collins, W. K. Michener, and M. A. Huston. 2008. "Living in an Increasingly Connected World: A Framework for Continental-Scale Environmental Science." *Frontiers in Ecology and the Environment* 6: 229–237.

Peterson, E. B. 2005. "Estimating Cover of an Invasive Grass (*Bromus tectorum*) Using Tobit Regression and Phenology Derived from Two Dates of Landsat ETM+ Data." *International Journal of Remote Sensing* 26: 2491–2507.

Poisot, T., A. Bruneau, A. Gonzalez, D. Gravel, and P. Peres-Neto. 2019. "Ecological Data Should Not Be So Hard to Find and Reuse." *Trends in Ecology & Evolution* 34: 494–96.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Record, S., M. A. Jarzyna, B. Hardiman, and A. D. Richardson. 2022. "Open Data Facilitate Resilience in Science During the COVID-19 Pandemic." *Frontiers in Ecology and the Environment* 20(2): 76–77. Portico. https://doi.org/10.1002/fee.2468.

Record, S., N. M. Voelker, P. L. Zarnetske, N. I. Wisnoski, J. D. Tonkin, C. Swan, L. Marazzi, et al. 2021. "Novel Insights to Be Gained From Applying Metacommunity Theory to Long-Term, Spatially Replicated Biodiversity Data." *Frontiers in Ecology and Evolution* 8, 612794. https://doi.org/10.3389/fevo.2020.612794

Richardson, A. D., K. Hufkens, T. Milliman, D. M. Aubrecht, M. Chen, J. M. Gray, M. R. Johnston, et al. 2018. "Tracking Vegetation Phenology across Diverse North American Biomes Using PhenoCam Imagery." *Scientific Data* 5: 180028.

Robinson-García, N., E. Jiménez-Contreras, and D. Torres-Salinas. 2016. "Analyzing Data Citation Practices Using the Data Citation Index." *Journal of the Association for Information Science and Technology* 67: 2964–75.

SanClements, M. D., S. Record, K. C. Rose, A. Donnelly, S. S. Chong, K. Duffy, A. Hallmark, et al. 2022. "People, Infrastructure, and Data: A Pathway to an Inclusive and Diverse Ecological Network of Networks." *Ecosphere* 13: e4262.

Schimel, D., W. Hargrove, F. Hoffman, and J. MacMahon. 2007. "NEON: A Hierarchically Designed National Ecological Network." *Frontiers in Ecology and the Environment* 5: 59.

Silvello, G. 2018. "Theory and Practice of Data Citation." *Journal of the Association for Information Science and Technology* 69: 6–20.

Slingsby, J. A., A. M. Wilson, B. Maitner, and G. R. Moncrieff. 2023. "Regional Ecological Forecasting across Scales: A Manifesto for a Biodiversity Hotspot." *Methods in Ecology and Evolution* 14: 757–770.

Soranno, P. A., L. C. Bacon, M. Beauchene, K. E. Bednar, E. G. Bissell, C. K. Boudreau, M. G. Boyer, et al. 2017. "LAGOS-NE: A Multi-Scaled Geospatial and Temporal Database of Lake Ecological Context and Water Quality for Thousands of US Lakes." *GigaScience* 6: gix101.

Stanley, E. H., L. C. Loken, N. J. Casson, S. K. Oliver, R. A. Sponseller, M. B. Wallin, L. Zhang, and G. Rocher-Ros. 2023. "GRiMeDB: The Global River Methane Database of Concentrations and Fluxes." *Earth System Science Data* 15: 2879–2926.

Stoudt, S., V. N. Vásquez, and C. C. Martinez. 2021. "Principles for Data Analysis Workflows." *PLoS Computational Biology* 17: e1008770.

Sturtevant, C., E. DeRego, S. Metzger, E. Ayres, D. Allen, T. Burlingame, N. Catolico, et al. 2022. "A Process Approach to Quality Management Doubles NEON Sensor Data Quality." *Methods in Ecology and Evolution* 13: 1849–65.

Thibault, K. M., C. M. Laney, K. M. Yule, N. M. Franz, and P. M. Mabee. 2023. "The US National Ecological Observatory Network and the Global Biodiversity Framework: National Research Infrastructure with a Global Reach." *Journal of Ecology and Environment* 47: 21. https://doi.org/10.5141/jee.23.076

Tobi, H., and J. K. Kampen. 2018. "Research Design: The Methodology for Interdisciplinary Research Framework." *Quality & Quantity* 52: 1209–25.

U.S. Geological Survey. 2023. *National Hydrography Dataset (NHD)*. U.S. Geological Survey, National Hydrography Dataset (ver. USGS National Hydrography Dataset Best Resolution (NHD) for Hydrologic Unit (HU) 4 - 2001 (published 20191002)). https://www.usgs.gov/national-hydrography/access-national-hydrography-products

Walther, S., S. Besnard, J. A. Nelson, T. S. El-Madany, M. Migliavacca, U. Weber, N. Carvalhais, et al. 2022. "Technical Note: A View From Space on Global Flux Towers by MODIS and Landsat: The FluxnetEO Data Set." *Biogeosciences* 19: 2805–40.

Wan, W., H. Li, H. Xie, Y. Hong, D. Long, L. Zhao, Z. Han, et al. 2017. "A Comprehensive Data Set of Lake Surface Water Temperature Over the Tibetan Plateau Derived From MODIS LST Products 2001–2015." *Scientific Data* 4: 170095.

Weinstein, B. G., S. Marconi, S. A. Bohlman, A. Zare, A. Singh, S. J. Graves, and E. P. White. 2021. "A Remote Sensing Derived Data Set of 100 Million Individual Tree Crowns for the National Ecological Observatory Network." *eLife* 10: e62922.

Welti, E. A. R., A. Joern, A. M. Ellison, D. C. Lightfoot, S. Record, N. Rodenhouse, E. H. Stanley, and M. Kaspari. 2021. "Studies of Insect Temporal Trends Must Account for the Complex Sampling Histories Inherent to Many Long-Term Monitoring Efforts." *Nature Ecology & Evolution* 5: 589–591.

White House Office of Science and Technology Policy (OSTP). 2022. "Desirable Characteristics of Data Repositories for Federally Funded Research." Executive Office of the President of the United States.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4: 1686.

Wieczorek, J., D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais. 2012. "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard." *PLoS One* 7: e29715.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 160018.

Wurzburger, N., K. J. Elliott, and C. F. Miniat. 2023. "Long-Term Changes in Forest Biomass, Tree Species Composition and Nitrogen Fixation Following Land Use Disturbance."

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Atkins, Jeff W., Kelly S. Aho, Xuan Chen, Andrew J. Elmore, Rich Fiorella, Wenqi Luo, Danica Lombardozzi, et al. 2025. "Recommendations for Developing, Documenting, and Distributing Data Products Derived from NEON Data." *Ecosphere* 16(1): e70159. https://doi.org/10.1002/ecs2.70159