Response to reviewers

01 February, 2022

15-Nov-2021

Dear Dr. Li:

Thank you very much for submitting your manuscript "Standardized NEON organismal data for biodiversity research" for review by Ecosphere. The reviewers and I appreciate the work you have accomplished. I am willing to consider a revised version for publication in the journal, assuming that you are able to modify the manuscript according to the recommendations. Your revisions should address the specific points made in the review comments, as well as my comments (below my signature).

To submit your revised manuscript, log into https://mc.manuscriptcentral.com/ecosphere and enter your Author Center. You will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision. Please DO NOT upload your revised manuscript as a new submission.

When submitting your revised manuscript, you will be able to respond to the review comments in the space provided. You can use this space to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the review comments. If you disagree with a review comment, please explain why. (Please note that the field does not retain type formats such as italics, boldface, or colors, so please format the responses accordingly.). Please include a "track changes" version of your file, along with a clean copy as your Main Document.

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please replace files of the earlier version and delete any redundant files before completing the submission.

Because the focus of Ecosphere is rapid publication, we request that you submit your revised manuscript within three weeks from today's date. If you are unable to meet this deadline, please request an extension by contacting the the journal's editorial staff at esajournals@esa.org. If you are unable to revise the manuscript within three months, you will need to resubmit the revision as a new manuscript.

Sincerely,

Dr. Jennifer Balch
Subject-matter Editor, Ecosphere
jennifer.balch@colorado.edu

Thank you for your help with our manuscript.

Comments from the Subject-matter Editor:

More specifically, I would appreciate if you can address these issues:

- Provide some guidelines on how to cite data that is obtained from the neonDivData package.
- Address the question about the longevity of the package and the intention to maintain it.
- Provide a brief discussion of the degrees of freedom issue in relation to the number of NEON sites.
- Given that this is a paper about NEON data it necessarily needs to be U.S. focused, but please describe if this effort is a model for using biodiversity data from other site locations and whether it is scalable? What might be some of the limitations? Or benefits of using this approach?
- Greater detail and description of the data and NEON sampling procedures and data cleaning (e.g., post-processing tests to detect errors), so that the information provided by the manuscript is accessible to a broader audience.

Dear Editor,

We have addressed all issues raised by reviewers. Please see our responses below and the revised manuscript for details. Again, thank you for your help with our manuscript.

Daijiang Li (on behalf of all co-authors)

Reviewer #1 (Comments to the Author):

Thank you for the opportunity to review the manuscript “Standardized NEON organismal data for biodiversity research”. The manuscript provides a well-written explanation of a new data package that provides NEON data in a standardized format for biodiversity and macroecological analyses. The manuscript includes brief explanations of the data cleaning and processing decisions that were made when preparing the standardized data. I think this manuscript and the data package are useful contributions. While the manuscript is long, I think the explanations of sampling design and data preparation for the different taxonomic group are useful, and justify the length. I think this manuscript is suitable for publication in Ecosphere with relatively minor changes, which are described below.

Thank you.

Detailed Comments:

Can you please include guidelines about how to cite data that is obtained from the neonDivData package? Should the citation be of the neonDivData package, or should users cite the NEON dataset or data contributors directly somehow? Perhaps you could include an example of how to cite the data used in your example analyses.

This is a great comment! We added a short guideline in the Results, which reads: "If this standardized version of NEON data was used, users should cite this paper along with the citations provided by NEON for each taxonomic group. Such citations can be found in the URLs presented in Table 1." We have also added a CITATION file in the `neonDivData` package so that R users can type `citation("neonDivData")` to get this information.

> How long into the future do you expect the data package will continue to be updated? In line 161 it says: "Standardized datasets will be maintained and updated as new data become available from the NEON portal." Who is doing this updating? How long do you expect to do this (a rough estimate is fine). Is this data package part of the NEON project, or is it an independent project, in terms of institutional and long-term support?

Sustainability is very important for projects like this. This is an independent project as a collaboration among scientists and NEON staff. However, we expect that our project will be updated at least in the next ten years. We are confident about this for a couple of reasons. First, we have added a GitHub Action workflow in the GitHub repository, which will automatically update all datasets at the beginning of March every year. We set the update date to March 1st because NEON usually releases new data in January each year. Second, all codes are openly available online and are organized into an automatic workflow. Therefore, it takes very little effort in the future to update the dataset manually if the GitHub Action Workflow failed. All three co-first authors agreed to maintain this project in the next several years.

> One question that I always think about with these types of datasets, and that the authors might be able to comment knowledgeably on, is about degrees of freedom. In one of the examples in this manuscript (line 720), the authors write that the example "... shows the utility of the data package for exploring macroecological patterns at the NEON site level." But if I understand correctly, there are only 81 NEON sites. This means a site-level analysis of macroecological patterns has a sample size of 81 (surveys at a site are pooled to the site level). With such a small sample size, degrees of freedom are a concern both within a single analysis (if multiple predictor variables area analyzed), and especially if many different researchers ask multiple questions of the same dataset. It seems to me that the strength of the NEON data is in the intensive and repeated sampling done within sites, not in having a large number of sites. Why do the authors think this dataset is well suited for analyses at the site level? My guess is that the data are better suited to study questions for which the repeated samples at sites all provide some partially independent information, and can all contribute some (partial) degrees of freedom. But I don't understand the NEON sampling scheme in detail, so maybe I am wrong about that. In any case, since the authors provide a data package that could potentially encourage very many analyses, I think the manuscript would be strengthened by a brief discussion of the degrees of freedom issue in relation to the number of NEON sites, and how much we as a research community can actually learn from analyzing a dataset like this again and again to answer different questions.

This is a good point and consideration of degrees of freedom is an important aspect of model fitting. We have toned down our language and the manuscript now reads as, "...show the utility of the data package for exploring macroecological patterns." We have left in our example of richness versus latitude as that only includes one predictor (latitude), and as the reviewer notes their concern mainly lies in analyses at the site level with more than one predictor. We feel that a thorough discussion of the various sample size considerations for analyses that could be performed with the NEON data or issues of Type I error associated with open source ecological data sets is beyond the scope of this manuscript, so we have not discussed this point in detail. However, we do agree that users of this data package should be well aware of considering sampling and sample size issues associated with the NEON data, so we have included a new reference to Barnett et al. (2019), which goes into great detail about various sample size issues associated with the NEON organismal data. Reference to this paper is now included in the last two sentences of the second paragraph of the Discussion section which reads as, "In a similar vein, researchers should ensure that they have considered sample size issues before fitting any models with these data. See Barnett et al. 2019 for a review of the NEON organismal sampling design that contains important insights related to sample size issues."

> Minor comments:
>
> Line 65: I think there is a word missing or the verb tense in the header is wrong. In "Introduction (or why standardized NEON organismal data)". Should that read "...why *we* standardized NEON...", or perhaps "...why standardize NEON..."?

Here, we treat 'standardized' as an adjective word.

> Fig. 1, panels B, C and D, the text on the figures is too small, and is hard to read unless I zoom my PDF viewer to 150% or more.

Done.

> Line 721: Add the word "rich" after "species", so it reads "...wherein sites are more species rich at lower latitudes..."

Done. Thanks!

Reviewer #2 (Comments to the Author):

> The work described in this paper is of extreme value. The authors have taken heterogeneous datasets from the NEON project and have converted them into a common structure and standard, with documented procedures and open-code. This has been done in collaboration with data collectors and NEON staff. This work will save hundreds of hours of work to other researchers while avoiding data misuse and will make comparisons between studies easier. It is simply an enormous and precious service provided to the scientific community! I definitively support this effort, but I think that the paper could reflect better this work by being more clear, better organised and more understandable for an international (non-NEON) audience.

Thank you for your constructive comments!

> 1. The paper is very US-focussed, and it sounds that NEON is unique in the world. However, there are examples of large and coordinated projects collecting observational data all over the world! It would be good to mention some of them, here are some examples (very far from being exhaustive and not necessarily the ones to cite, but just to give some examples): Fundiv Europe (http://project.fundiveurope.eu/), Biodiversity Exploratories (https://www.biodiversity-exploratories.de/en/), UK National Biodiversity network (https://nbn.org.uk/tools-and-resources/useful-websites/database-of-wildlife-surveys-and-recording-schemes/), Brazilian Network of Networks (https://www.sciencedirect.com/science/article/pii/S2530064418300336), and surely many others.

This is an excellent suggestion for the paper, and we agree that mentioning these other networks will increase the paper's relevance to the broader global community. To this end, we have added this statement to the last sentence of the abstract, "...and encourage the use of the harmonized ecocomDP data design pattern for community ecology data from other ecological observatory networks." We have also added this sentence to the very end of the paper in the Conclusion section, "Finally, extension of the ecocomDP harmonized data design pattern to data from other ecological research and observatory networks (e.g., the Brazilian Network of Networks [Oliveira Roque et al. 2018], South African Environment Observation Network [Jaarsveld et al. 2007]) has the potential to enable community ecologists to better synthesize data from across the globe."

2. As it is, the paper targets an audience of readers that are familiar with NEON and data (re-)use in NEON. Maybe this is fine as I see that the paper is part of a "special feature". But be aware that someone unfamiliar with NEON (like myself!) cannot really understand the paper, and this makes it look more like an "internal" manual than a paper for an international audience. The introduction goes from very general statements to a very short description of NEON (L.117) and then to very technical data issues, which is hard to understand without context. I generally missed a more detailed description of the data situation and procedures in NEON: how many datasets there are, what the data requirements are (e.g. what is the minimal information that is required when uploading a dataset), what data we are speaking about (occurrences/abundances/traits?), who collects the data, what is the data flow, how data is checked, where it is stored... It was hard for me to imagine how the data looks like pre-cleaning: is there a dataset per taxon and site and year? Or is there one dataset per taxa? Are these static or data is added to a dataset every year? What is the shape of the datasets, are they in the long format or are they siteXspecies matrices? Is this homogeneous or heterogeneous across taxa? What is standardised and what is not? This might be detailed in other papers (or obvious when accessing the data repository), but it is important to be described here (in the introduction or methods) in order to understand the rest of the text.

We hear you regarding all your questions here. In fact, we felt the same way when we started working on NEON data, which partly motivated us to have a standardized version of NEON organismal data to facilitate biodiversity research using NEON data. For each taxon, NEON's raw dataset includes multiple tables in long formats. Because of the complexity, different taxons do not necessarily have the same data structure. It also makes it very hard to describe all the details for each taxon here in this already long manuscript. Furthermore, we think it is beyond the scope of this manuscript to introduce and describe the detailed protocol of NEON. Readers who want to learn more details about each taxon should check out the links presented in Table 1. These links were also included in the R package documentations for the data for each taxon.

3. Material and methods first section - The material and methods have a general section and then go into the specifics for each taxon. What I missed here was a section providing the general principles of data cleaning. When doing such data cleaning there is always a balance between providing a larger dataset with many possibilities of analyses but with potential misuse/mistakes or providing a smaller/less flexible but "safer" dataset. What was the choice here? How did the authors proceed in general (keep more data / keep only robust data)? Are there any general decisions that were applied to all taxa? What comes to my mind is (in a random order):

- was there a general protocol to be applied to all taxa?

Because different taxons have different sampling designs, there is no general protocol to be applied to all taxa.

- what is the minimal information needed for all taxa? (this can be guessed while reading the text but it is not explicitly described)

The minimal information needed for biodiversity study for all taxa includes locations (`location_id`, `site_id`, `plot_id`, etc.), species names (`taxon_id`, `taxon_rank`, etc.), and presence/absence or abundance information.

- what is the final output? A dataset per taxon and all years, that can be easily merged (or even "rbinded") with other similar datasets for other taxa? One dataset per year and taxon?

The final output for each taxon is a data frame with all data accumulated to the latest NEON release. It is possible to merge all taxa data together. However, instead of merging raw observation data, it is probably better to do this after calculating biodiversity values at each site.

- general strategy to deal with NAs?

NAs are removed for all taxa except plants, which have both presence/absence and abundance information (at 1 m^2 scale) and NAs are used when no abundance information was available.

- general strategy to deal with zeros? Are zeros reported? If so, for example if species A is detected in 2016 but not in 2017, is it reported in the 2017 dataset with a 0?

Yes, we did report 0s for all taxa.

- traits (e.g. body size) are mentioned for some groups, are these included in the cleaned datasets or is the focus only on occurrences/abundances?

Traits are not collected for most taxa so the main focus here is on occurrences and abundances; but we did include traits information whenever available from the NEON portal.

- how to deal with different sampling efforts for different sites (e.g. l.205 "5 to 25 points may be surveyed" or L.377 plants are surveyed once or twice per year -> how is this dealt with in general?)

Standardizing sampling efforts is a major component of this project and we have done this for all taxa. For each taxon, the `value` column includes standardized abundance information.

> Note that some paragraphs in the discussion mention these general points and would be better suited in the methods section (e.g. L.774, L.818 and L.837 which would be a nice first methods paragraph actually!)

We have moved this paragraph into the beginning of the Methods section.

4. Material and methods taxa - This detailed description is certainly needed but I found it was the hardest part to follow. It is difficult to get the specifics without knowing the original datasets. My suggestion would be to start with a description of the minimal columns in a dataset, those that are there for all taxa and could potentially be used to create a multi-taxa dataset. Then discuss the differences and explain why there are extra / different columns in different datasets. This would help to homogenise

the sections for each taxon with some common information. A possibility would be to provide the input and output headers of each dataset with short descriptions, maybe in a large table, maybe in the appendix. If a common vocabulary is applied to all taxa, it would also be helpful to describe what it means for each of taxon. For instance, "bout" is used for several taxa but (obviously) corresponds to different things for beetles or birds. It would be nice to have all these definitions in a single table or in a standardised way in each section.

As suggested by the reviewer, we have added an Appendix to print out the headers of each data product, and to describe the meaning of each column in these data products. This information is also available in the package website (https://daijiang.github.io/neonDivData/index.html) and the EDI repository (link here).

5. There is no description for systematic post-processing tests to detect potential mistakes in code or potential wrong decisions. Did the authors consider this? For instance, possibilities would be to count NAs or number of taxa in original and cleaned datasets. Or perform simple correlations, e.g. calculate richness correlations between original and cleaned datasets.

We take our data cleaning process seriously even though we did not put any description in the main text about it. We made our decisions for each taxon group after several group discussions. Our code to process each taxon was also internally cross reviewed by other sub-groups. After then, they were integrated into the R package `ecocomDP` by one of the leading authors (Eric Sokol, data scientists at NEON), who also did a final review of all codes. Furthermore, we have added test codes in the R package to test for potential obvious mistakes. Therefore, we are relatively confident that our codes match our decisions for data processing. Because we mostly just remove information that was not needed for general biodiversity studies, we don't think it would be very helpful to calculate correlations between richness values based on original and cleaned datasets.

6. The discussion still contains descriptions of methodological decisions. I think that these would fit better into the methods section (see above). What I missed was a more general perspective on common issues with data from large projects and how to prevent them in future projects.

We have moved such parts into the Methods section. Each project, especially large ones, has its own issues. Talking about these unique as well as common issues is beyond the scope of this already long manuscript (but it can be a great perspective paper in the future). Therefore, we did not get to this in the revision.

7. R package documentation - The authors have to assume that users won't dig into the code and won't necessarily check the original datasets so accurate descriptions of changes in the datasets are very important, these include:

- be consistent in different sections, especially in the "description" and "details" sections. For instance the "details" section for mosquitoes is informative while the one for birds is vague and the one for fishes (and several other taxa) is absent. The description section for the fish survey contains what is in the "details" section for other taxa.
- Please specify any information that has been removed from the original dataset. For instance, the description of the bird dataset states: "We only removed some columns that likely won't be used in biodiversity studies." This can be a big assumption! Please

at least report the names of these removed columns so users know they exist (e.g. L.225 “observer name” might actually be an interesting variable to correct for potential bias).

- Please specify large limitations or very important things to know for each dataset (e.g what is described L.256)

We have updated the package documentations according to the suggestions. Thank you!

> Minor comments: - L.130: “units must be calculated”: doesn’t sound correct, maybe rephrase

We have replaced ‘calculated’ with ‘decided’.

- L.131: metadata unnecessary: this is a dangerous approach, we never know what will be necessary in future. Maybe provide an example.

We have removed this part of the sentence.

- L. 154 and elsewhere: organismal data: please define what this is: occurrences, abundances, traits?

We added ‘(e.g., occurrence and abundance of species)’ after the first appearance of organismal data in the Introduction.

- L. 225: I would have actually kept “observer name” to look at potential observers bias

As noted by the reviewer, it is hard to find a good balance between a detailed database with lots of flexibility and a cleaner one with less flexibility. We decided to remove the “observed name” in the final data product. If users want to get such information, they can easily change our open source raw code to retain it.

- L. 239: how is this handled? For a temporal analysis I would remove North pitfall traps for all years.

There is a column named `trapID` in the `data_beetle`. Users can filter out those North pitfall traps (`trapID == "N"`) for years before 2018 if they need to remove those traps.

- L. 721: “wherein sites are more species at lower latitudes”-> more species rich maybe

Added.

- Please shortly explain what “EDI” is

EDI stands for Environmental Data Initiative and was explained in the Introduction when we first introduced it.

- The code in the supplementary could be amended with the code to produce the actual figure as such maps are likely to be produced when using NEON data and for the beetles analysis, it seems a bit odd to only have a small chunk of the code available.

All codes to generate figures are available in the GitHub repository (https://github.com/daijiang/neonDivData/blob/master/manuscript/figures.R). We have also added them in the Appendix.