Biogeosciences Discussions

Biogeosciences Discuss., author comment AC1
https://doi.org/10.5194/bg-2021-323-AC1, 2022

Reply on RC1

Katherine E. O. Todd-Brown et al.

Author comment on "Reviews and syntheses: The promise of big diverse soil data, moving current practices towards future potential" by Katherine E. O. Todd-Brown et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2021-323-AC1, 2022

We would like to thank the reviewer for their comments and note that the original review is in *italic* with our reply following.

General comments

This narrative/opinion manuscript describes the promise and challenges of soil databases, data discovery and harmonization, and related efforts. This is a topical and timely topic, given the many top-down and bottom-up efforts that have sprung up in this area over the last 10-15 years. The manuscript is reasonably well written and has interesting points, but I think there are some significant weaknesses here as well.

*First, the authors seem to repeatedly conflate and/or mix up "big" and "open" data—starting with the title, see #1 below. They're obviously not the same thing, and most of the ms seems to actually focus on *open* data. It would be good to clearly define these terms, their distinctions, and use the terminology consistently and correctly throughout.*

The authors respect that there could be some confusion for readers when it comes to how we are referring to "Big Data" within the manuscript. We believe that 'big' extends beyond the absolute size of the data files and includes 'diverse' data as well; we now pair these two words consistently throughout the paper, including in the title. We also disagree that our manuscript focuses on "open" data, although it is understandable that the reviewer came away with this impression. All of the strategies discussed in this paper could be applied to 'closed' or proprietary data, although we acknowledge additional difficulties with proprietary data in section 3.1. We suggest the following changes that will hopefully help clarify this point.

- A new title: "The promise of big diverse soil data…"
- Abstract ln 2: "In the age of big data, soil data are more available and richer than ever…"
- Ln 50: It's important to note here that FAIR does not always mean open freely reusable data. Indeed the FAIR Data motto makes this difference quite clear: "as open as possible, as closed as necessary", and this becomes particularly important for data that has possible economic impact (Luque 2019).

- (Luque, C., 2019. Open data and FAIR data: Differences and similarities. Plataforma OGoov [en ligne], 23.; "Open Data". (2019), OGoov Open Government Platform, 6 May, available at: https://www.ogoov.com/en/glosario/open-data/ (accessed 6 March 2022).; and "la Investigación", B.U.A. al A. y. (2017), "Biblioguías: Datos de investigación: gestión, datos abiertos (Open Data): INTRODUCCION", available at: https://biblioguias.unex.es/datos_de_investigacion (accessed 6 March 2022).)

- Ln 183: "To reach this hope it's important not just to work with large volumes of data but also diverse observation and measurements, and do so in a way that is trusted."

Thank you for pointing this out to us and giving us an opportunity to improve the paper.

Second, what about this is unique to soils? I struggled to find anything in the ms that wouldn't apply to environmental data more generally, and thus what points are made here that haven't already been made by authors like Wolkovich 2012 (http://dx.doi.org/10.1111/j.1365-2486.2012.02693.x) or Rüegg et al. 2014 (http://dx.doi.org/10.1890/120375). What exactly is the value added here, in other words? That's not as clear as it needs to be.

Both reviewers bring up this point that soils data needs are not unique. And we agree that the strategies for creating and maintaining soil databases are not unique. We have modified the introduction and conclusion to highlight the relevance of these findings to the broader environmental community. Soils are unique in their importance and societal relevance but not in these identified data challenges. This paper strives to lay out how soil scientists currently aggregate data and point out various strengths and weaknesses of this approach. It is, by design, experiential and written by soil scientists not informaticians (see lines 59-61). Specifically we proposed adding on ln 60: "The approach and issues outlined in this paper are undoubtedly not unique to soils and are relevant to a wide range of scientific data, particularly environmental data. However we present this as a case study of soil specific database construction."

Wolkovich etal 2012 contends that the problem is motivation and knowledge on the part of the data collectors/providers; which we disagree with, data providers are often highly motivated to see their data have as broad and large an impact as possible. We will add the following statement in the introduction on line 50. "Indeed previous research has identified challenges with educating and motivating data providers to publish their data sets (Wolkovich et al 2012)."

Ruegg etal 2014 does come to similar conclusions that a common descriptive framework would benefit the field more broadly. However we show how researchers are currently conducting data harmonization without such resources in this paper and that folks will generally find this template approach both too large and too small. In addition, we would contend that their suggestion of informatics being an early part of the project design is not tractable for most small, single PI data and demonstrate with this manuscript that researchers are moving the field forward without this element. We will add Ruegg etal 2014 as an example of a suggested standard approach on line 240.

Third, there are many curious omissions from the references, I thought. For example, Crystal-Ornelas et al. 2021 ("A guide to using GitHub for developing and versioning data standards and reporting formats", Earth Space Sci.,

https://doi.org/10.1029/2021ea001797) is relevant in many places.

Thank you for bringing Crystal-Ornelas 2021 to our attention, we've added this to the recommended best practices for community semantic tool development. Specifically we propose adding to ln 245: "...practices from the open source community such as version control also are a critical tool (Crystal-Ornelas et al 2021)"

Re reporting formats, Bond-Lamberty et al. 2021 ("A reporting format for field measurements of soil respiration", Ecol. Inform., 62, 101280), which was part of a special issue on integrating long-tail data: https://www.sciencedirect.com/journal/ecological-informatics/special-issue/101T38RSLSF.

Again, thank you for bringing this 2021 paper to our attention and we have added it as a contrasting example of data standards on line 239: "...(for example, the format suggested by Bond-Lamberty et al 2021"). However we will point out that one of our main findings was that data standards as described here are insufficient due to the diversity of measurements and study design.

In a related vein, the SRDB (https://github.com/bpbond/srdb) is a decade older than most of the efforts discussed here and widely used and cited, so might be worth a mention as well, unless you're particularly focusing on stocks but not fluxes.

We agree that SRDB is an excellent example of soil data harmonization, and further add that the Worldwide soil carbon and nitrogen data Zinke et al 1986 is an even older example of soil data harmonization. We've added this to our introduction and included a table of active/recent soil database projects. We are removing the ILAMB sections (ln 41-45) and replacing this with the following: "A number of databases have been compiled in soils data around specific themes or measurement types including: soil carbon and nitrogen (Worldwide soil carbon and nitrogen data Zinke et al 1986; International Soil Carbon Network database ISCN Nave 2015), field based soil respiration (Soil Respiration Data base; Bond-Lamberty and Thomson 2010, Jian et al 2021), lab-based heterotrophic respiration (Soil Incubation Database), soil radiocarbon (International Soil Radiocarbon Database), and coastal soils (Coastal Carbon Research Coordination Network Database) (See Table XX for a complete list with database properties)."

Unlike most of the other efforts discussed, SoilGrids (Hengl et al.) really is big data (pretty big anyway) and that should be noted.

SoilGrids is an excellent collection of data products that highlights how we differentiate between databases and data products in this paper. We address this, and related comments from R2, beginning on line 45.

Suggested text:

Soil resources curated by ISRIC (https://www.isric.org/) provide another example of how soil data feed into larger products. After archival on ISRIC servers, datasets from individual providers are incorporated into the World Soil Information Service workflow (WoSIS; https://www.isric.org/explore/wosis). The WoSIS workflow includes mapping diverse data contributions to a standard data model, harmonization, and distribution. Distribution includes a database, as defined in this paper (the WoSIS Soil Profile Database; https://www.isric.org/explore/wosis/faq-wosis#How_should_the_WoSIS_datasets_be_cited?), as well as derived data products, such as SoilGrids (Hengl T, de Jesus JM, MacMillan RA, Batjes NH, Heuvelink GBM, et al. (2014) SoilGrids1km — Global Soil Information Based on Automated Mapping. PLoS ONE 9(8): e105992. doi:10.1371/journal.pone.0105992)

In summary, there are many points of interest here, and I applaud this effort by the authors. The current ms has some significant issues, and would benefit from tighter language—it's pretty long—and clearer novelty.

We highlighted the generality and novelty of the soil centered nature of this paper in the above change to the introduction. We tightened the language, for example, by removing the ILAMB reference in the introduction and replacing it with a review of a selection of current soils databases. We hope that this addressed your concerns.

Specific comments

Most of these were specifically addressed above. However for completeness we include the following responses.

- *Title: a bit odd (most of this manuscript is about *open* data, not *big* data), and it's a run-on sentence; consider rewording*

We hope that our clarification in the introduction addressed this wording and suggest revising the title to include "big diverse soil data".

- *Line 182: do you mean "open" data here? That's not what big data is*

See above.

- *193: …just like any other environmental data*

See above.

- *212: wow, that (60%) is shocking*

We know right?!?!

- *215-: do you mean "time" of collection, i.e. 1400 hours? Or "date"?*

Good point. We suggest adding the following to ln 217: "While the exact resolution will vary depending on the measurement (for example, parent material may just need the decade of collection while soil respiration may need a minute resolution), recording [...]"

- *239: see recent ESS-DIVE -funded papers on data standards/reporting formats in Ecological Informatics*

See above.

- *290: a better analogy might be the *software* review process? See Crystal-Ornelas paper*

See above.

- *296-312: this is all restating material above, should be removed*

We respectfully feel that this ties the ending of the paper back to the introduction and suggest keeping this section.

- *403: haha, data, singular or plural? Both!*

Clearly! Thank you for the catch. We've replaced this with "data are"

- *433: what is this referencing? Confusing*

We've added the following leading sentence for this paragraph on ln 433: Data privacy concerns and the impact of 'good/bad' data metrics on land valuation are still an issue but "trusted' data holders are attempting to address this.