Biogeosciences Discussions

Biogeosciences Discuss., referee comment RC1
https://doi.org/10.5194/bg-2021-323-RC1, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on bg-2021-323

Anonymous Referee #1

Referee 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-RC1, 2022

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.

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.

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

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.

Specific comments

- Title: a bit odd (most of this manuscript is about *open* data, not *big* data), and it's a run-on sentence; consider rewording
- Line 182: do you mean "open" data here? That's not what big data is
- 193: …just like any other environmental data
- 212: wow, that (60%) is shocking
- 215-: do you mean "time" of collection, i.e. 1400 hours? Or "date"?
- 239: see recent ESS-DIVE -funded papers on data standards/reporting formats in Ecological Informatics
- 290: a better analogy might be the *software* review process? See Crystal-Ornelas paper
- 296-312: this is all restating material above, should be removed
- 403: haha, data, singular or plural? Both!
- 433: what is this referencing? Confusing