Comments from Jeffrey Beem Miller, Referee #1, are provided below in normal text.
Our responses to each are below each comment in bold with suggested changes to the revised manuscript identified by text in quotes. Note, this is easier to see in the .pdf files we've uploaded along with this plain text response.

General comments

The authors present a new database (SoDaH) of soil observations synthesized from datasets curated by five well-known research networks: LTER, CZO, NEON, NutNet, and DIRT. Two key aspects of SoDaH make it unique among the new generation of soil databases: 1) flexibility in its approach to harmonizing data from diverse sources, and 2) the development of a powerful web-based tool for querying, visualizing, and extracting subsets of data. These two features of SoDaH make it a valuable addition to the growing pool of soil databases available to the soil science community.
Thank you, we appreciate recognition of these novel aspects of the database.

The manuscript is generally well written, and clearly presents the need for more accessible sources of compatible soil data to facilitate broad-scale syntheses, as well as the challenges of providing such a resource. However, the description of the process of data harmonization and aggregation is somewhat confusing in the text (Fig. 3 provides an excellent visual summary). Please see the specific comments for more details. Additionally, the metadata template only provides sparse instructions for how to properly fill it out. I had to carefully compare a few of the source data files with the accompanying (filled out) metadata template in order to understand exactly what data was required for the metadata template and in what format it should be provided. A more thorough guide or an additional supporting document would improve this process. For example, the ISRaD database (Lawrence et al., 2020) mentioned in this manuscript provides both a template file and a supporting "template information file" to facilitate data entry. I would recommend implementing a solution along these lines.
This echoes comments made by R2. We have modified the website to include a new 'contribute data' tab (https://lter.github.io/som-website/contribute_data.html) that includes:

- **Instructions provided here give an overview of how to contribute data.**
- **This link provides access to the SoDaH database template file**
    - ***Note that you will need to copy the database template file to be able to edit it.***
- **An example google directory, with primary data and a completed metadata template is provided for reference.**

One important feature of a synthesized database is data transparency, which SoDaH excels at in some ways, but falls short in others. The workflow for aggregating data preserves the raw data, which is the gold standard. However, with the way the metadata template is structured there is no clear way to document the data source for site level data, which data contributors enter manually. As multiple contributors could provide data from the same site, it seems

possible that conflicting data could be reported for, say, mean annual temperature. How would users distinguish which reported value is more appropriate for their analysis?
This comment is 100% accurate. Any synthesis is only as good as the data that is contributed to it. Manually entering data into the metadata template does introduce potential sources of error, but by creating a scripted infrastructure to generate the harmonized level-2 data we can go back to correct errors that may have been introduced with the level-0 contributions. As such, we hope users of the database will let the steering committee know when they find discrepancies or inconsistencies in the data.

Finally, a critical feature of SoDaH is the web-based tool for querying and generating reports from the database (Shiny app), but unfortunately the use of this tool is neither well documented in the text, nor by the supporting resources (with the exception of an hour-long webinar available as a downloadable video clip, which while very useful, is not very user-friendly). Underselling this extremely powerful feature of SoDaH is in my view the biggest shortcoming in SoDaH as presented here. Providing a simplified overview or vignette that gives an example of the kinds of queries that can be made (filters, etc.) and the reports that can be generated (visualizations, downloadable .csv tables, maps, etc.) would greatly help with reception and use of SoDaH within the community.
Thank you for the care in looking at the work we've contributed as part of SoDaH. We agree that the shiny app is really powerful and suggest the following text to help guide users through its functionality

"To facilitate user interaction with the SoDaH database, and to provide a simplified approach for data queries and analysis, we developed a web-based application using R Shiny (Chang et al. 2020). This SoDaH application is publicly accessible and hosted by the National Center for Ecological Analysis and Synthesis (NCEAS) at https://cosima.nceas.ucsb.edu/lter-som (last accessed July 15, 2020; source code: https://github.com/lter/lterwg-som-shiny). With the SoDaH application, users can perform a number of tasks to aid data discovery, visualization and analysis. We provide a brief description of this resource that highlights key features of the R Shiny SoDaH application.

In the *Query* section of the application, the top portion of the page provides a variety of data filter options to assist users with partitioning the database. Specifically, users may subset the database by any combination of research network, experiment type, and soil depth, while also specifying whether they wish to include or exclude experimental treatments or time-series data. Below the filter options, the *Output* section of the page contains three separate features arranged into labeled application tabs. The Plot tab allows users to quickly create basic analysis plots (point, histogram, or boxplot) using both covariates (e.g., Fe concentration) and metadata (e.g., mean annual precipitation). In the *Map* tab, users may specify which analyte in the database to display on a spatial map. Numeric values are symbolized using a color gradient and the interactive map functionality allows users to both adjust the map scale and select from numerous basemap options. Finally, the *Table* tab provides users with the ability to directly view, search and download the user-specified data subset as a flat file (.csv). The

plot, map and table features are all responsive to user specified changes in the data filters and will update in realtime.

In the *Data Summary* section of the SoDaH application, two feature tabs are provided to help users identify the data available for a specific site or analyte. The *By Analytes* tab allows users to view the number of analyte values that exist across all of the unique sites in the database. Users may specify up to four different analytes at a time to be included in the summary table output. The *By Site* tab allows users to view all of the analyte data available for a specific site. As the amount of data may be quite large for some sites, options are provided to narrow the summary output to include only profile, location or character class data.

The SoDaH application also includes a *Data Key* section, where users may view a full copy of the metadata template used for the SoDaH database construction, including descriptions of database fields and their associated metadata. The searchable key is split into two sections, location and profile, in the same manner as the metadata template used to describe raw data for the harmonization process. Field names in the provided key match exactly with analyte and metadata options provided in the *Plot* and *Map* features in the *Query* section of the application. Finally, the application provides a *Comments* section where users may submit an inquiry about the database or the application."

With minor improvements to the clarity of the text, and some additional documentation of the usage of the Shiny app, I think the manuscript is an excellent candidate for publication in ESSD.
Thank you for this positive assessment

Specific comments

Lines 58-60: Terms such as "harmonize" and "automate" would benefit from explicit definitions, although understandably this may not be possible with the word limits of the abstract (perhaps in the main text?).
We removed these phrases from the abstract. In the main text 'automated' is replaced with 'scripted', which is more accurate. We also clarify that "In the harmonized dataset, we convert analyte names and units to a standard output"

Line 83: It would be helpful to expand on or quantify exactly what you mean by "similar data products". Expanding on the importance of public availability of these databases would also be helpful.
We now define these similar data *syntheses*. "Providing similar data syntheses with information on soil carbon and associated covariates (e.g., climate, productivity, and soil physical and chemical properties) in public databases is critical to advancing understanding soil biogeochemistry."

Lines 123-130: Fine as is, but perhaps this information could be simplified in the text and expanded on in a table?

Given the data we're actually collecting are well documented on the website, metadata template, and Shiny App we'll avoid provided in table here that may give readers an abbreviated understanding of the data SoDaH contains.

Line 145: While the framework for reporting data from experimental manipulations is a key asset of SoDaH, it is not clear why SoDaH allows for a greater range of spatial and temporal data than other databases.
Reviewer #2 raised similar concerns. We've now clarified that "data from these kinds of studies should be incorporated into existing database structures, like ISCN, but the additional metadata requested as part of SoDaH helps database users understand more information about how data were collected from individual studies."

Line 168: Suggest moving "Fig. 3" to the end of the sentence (after "...the structure of SoDaH")
Done

Line 180: The use of "ontologies" in this context is not entirely unclear and sounds like jargon. Additionally, what does "automatic harvesting of data" mean? Do you mean something more along the lines of automating the process of data acquisition? I realize those are similar, but the meaning is not clear from how it is written.
We suggest replacing 'automated' with 'scripted' here and clarify "Ultimately, sophisticated metadata, such as controlled vocabularies and other, more expressive semantic technologies, may facilitate scripted harvesting of data from disparate networks and repositories (e.g., see review by Buck et al. 2019 for trends and examples in Marine Science).

Lines 182-185: It seems like the "site" (or "location"?) is the fundamental organizational unit of SoDaH. It might be helpful to state that more clearly and expand on the example provided in order to help readers understand how to define a site/location and how that definition relates to data organization at each level of SoDaH, i.e. from raw data to querying the aggregated database.
Reviewer #2 raised similar questions: We suggest expanding this paragraph as follows: The metadata template in SoDaH matches site-level information with the detailed measurements collected at each study site. Data on the location tab represents site characteristics for a single site or location (e.g., Prospect Hill Warming experiment at Harvard Forest). Accordingly, the harmonization script broadcasts data provided on the location tab (latitude, longitude, mean annual temperature, etc.) to every row of the harmonized dataset. Data on the profile tab includes profile information about experimental levels (e.g., plots within experimental blocks) and experimental treatments (e.g. +N fertilization) that help clarify how the data were collected. Data on the profile tab should also correspond to columns of variables that are reported in the Level-0 data (e.g., soil organic C measured at different soil layers). Accordingly, the harmonization script copies each unique measurement from the profile tab into a column of data in the harmonized dataset. Data contributors, therefore, can move variables from the location to profile tabs when appropriate. For example, NutNet and NEON data were submitted to

SoDaH with information from multiple sites on a single .csv file that provided information about each site as unique columns of data. We, therefore, moved site information (e.g., climate, latitude and longitude) onto the profile tab for these networks. Similarly, gradient studies that report tabular data for individual soil profiles can move information on slope, aspect, vegetation communities or parent material (typically on the location tab) onto the profile tab of the metadata template.

Lines 186-189: This section is not clear to me. What do you mean by the statement that data on the location tab are “broadcast to every row of the harmonized dataset”? The analogy that clarified this somewhat for me (used elsewhere in the manuscript) is the idea that the profile tab is a “map” for matching variables in the raw data onto the standardized variables in SoDaH. Additionally it is not clear why or when it would be appropriate to move data from the site to the profile tab.
Please see response to the previous comment, above.

Line 194: Can you expand upon (either in the text or by providing an example in the supporting information) how one would go about describing additional aggregation steps and how that would be implemented in the data aggregation process? This seems like a very messy and case-by-case basis, but also like a problem that would be encountered fairly frequently.
We’re still working on a robust way to do this, but will code to the github repository. However, because SoDaH is a flat database values from these different data files will be stacked, meaning that information from different Level-0 datasets would be recorded in different rows of the aggregated Level-2 database (in the example above, soil properties and productivity will be included, but in different rows). Additional aggregation steps, therefore, may be required to align data within sites. Users can find this information in the database column labeled merge_align, which is a logical that identifies if multiple data files can be merged. Notes under columns align_1 and align_2 are intended to help communicate what common data fields can help with this alignment (e.g. experimental or treatment levels, L1 and tx_L1, respectively). To help users understand the database column information, the complete database key is provided in the SoDaH online application and gives users descriptions of the column contents.

We also note in section 3.3: “As mentioned in section 2.4, future contributions of code to analyse the SoDaH database are encouraged. These contributions should be made to the LTER SOM GitHub repository, with a priority on developing additional utilities to align and aggregate datasets from individual sites and locations. Contributions will be reviewed by the SoDaH steering committee (currently Wieder, Pierson and Earl) and made publicly available. The committee will continue oversight while new funding options and/or partnerships (e.g., ISCN) are explored.

Line 200: When and how (what platform) would users “point to the Google Drive directory”? I assume this means when running the function in R?

This is clarified in section 2.3 "We developed the *soilHarmonization* package in R (R Core Team 2020) to harmonize and aggregate the SoDaH database"

Line 201: Suggest "...generates a new flat file(s) in which the relevant variable names and units are standardized..."
Done

Line 202-206: If possible, it would be helpful to define or clarify some of the terms you use throughout this section in advance, e.g. "harmonized dataset", "Level-1" data products". The process is very clearly shown in Fig. 3 along with the terminology, so perhaps you could give a one-sentence description of the workflow in which you name the outputs of each step of the process?
This may be appropriate as the last paragraph of section 2.1. The workflow for synthesizing is summarized in Figure 3 and in the following sections. Briefly, Primary data (Level-0) are identified by data providers and variables are mapped to standardized units and vocabulary using the metadata templates (section 2.2). These data are harmonized into Level-1 data with soil harmonization script that renames variables, conducts unit conversions, and performs quality control checks (section 2.3). Finally, Level-1 data are aggregated into the Level-2 dataset, which can be visualized with the SoDah R Shiny app and queried with data analysis tools (section 2.4).

Lines 208-212: This may not be the best place for it, but some discussion of data transparency would help to showcase the strengths SoDaH. Aside from the issue of C4 site-level data lacking a clear source, the preservation of raw data in SoDaH is a valuable feature. However, it is not clear from the website how to access the raw data files (I was able to find them, but it wasn't simple). Perhaps this could be clarified or stated explicitly somewhere?
In section 2.2 we note: "These primary data may or may not be in a published state but, if not published, would be equivalent to data provided for publication. Many of the datasets in SoDaH were already published in public repositories like EDI, the repository for LTER data, or available through the NEON data portal. Users can find these primary data using the doi provided for the individual dataset in the harmonized dataset. Other datasets that we wanted to include in SoDaH, however, had not been published or were difficult to find or identify (mainly data from CZO sites and the DIRT network, but also some LTER data). Publishing these primary data remains an active priority for our working group."

From a reproducibility standpoint, we probably should be storing the completed metatemplates in Zenodo, or add them to the current database repository in EDI. We wonder, however, how much value would be gained from such an effort?

Line 239: Clarify in the template how to specify these grouping variables.
In section 2.2 we note "Additional aggregation steps, therefore, may be required to align data within sites. Users can find this information in the database column labeled merge_align, which is a logical that identifies if multiple data files can be merged. Notes under columns align_1 and align_2 are intended to help communicate what common data

fields can help with this alignment (e.g. experimental or treatment levels, L1 and tx_L1, respectively)." and reference this section here in the text.

Lines 251-255: Perhaps histograms of these data could replace Fig. 5? In its current format Fig. 5 is completely illegible.
"Our intent with this figure is to illustrate the number of sites in each network, the temporal length of their data record, and the depth to which soils are typically sampled" With respect, we'd prefer the figure as-is to illustrate these points and clarify the intent in the figure caption

Table 1: Consider providing some examples of gradient studies or time series.
This seems to make the most sense in the table heading "Gradient studies may include measurements along a hillslope catena (e.g., several CZO sites), across vegetation communities (typically LTER sites), or surveys intended to capture local- to regional-variability (especially NEON periodic soil sampling). Time series studies involve repeated measurements in the same sites over time (LTER and NEON) and they which may also include experimental manipulations (e.g., NutNet, DIRT, & LTER)."

Figure 3: Excellent figure, very clearly describes the data workflow.
Thank you

Figure 5: This figure is illegible and it is unclear what it shows. Suggest replacing with histograms of site characteristics (see comment for lines 251-255).
See also our response, above.