Earth Syst. Sci. Data Discuss., referee comment RC3
https://doi.org/10.5194/essd-2021-454-RC3, 2022

Comment on essd-2021-454

Anonymous Referee #3

Referee comment on "A compiled soil respiration dataset at different time scales for forest ecosystems across China from 2000 to 2018" by Hongru Sun et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-454-RC3, 2022

Based on a thorough review of 568 original research articles and other publications, Sun et al. compiled a comprehensive soil respiration dataset that covers a wide range of climates, elevations and forest ecosystems across China. The dataset comprises a total of 10288 monthly and 634 annual soil respiration measurements and for some sites also monthly soil temperature measurements (at 5 and/or 10 cm depth). In addition, specific information (geographic location, forest type, mean annual air temperature and precipitation, etc.) are provided for each site. In view of the vast number of independent soil respiration studies from different regions, consistent datasets that summarise the state of the art and present the available data in a common format are of great benefit for the research community as they facilitate, for example, the analysis of spatial variations and temporal trends in soil carbon emission. The manuscript is generally well-written and the entire dataset is made publicly available through the open-access data repository PANGAEA. However, I have some concerns and suggestions that I would like to see addressed before I can recommend the manuscript for publication.

As I am not an expert in the field of soil respiration, my comments focus mainly on the overall content of the manuscript and the structure of the dataset.

General comments

Uniqueness of the dataset: A similar dataset as the one presented by Sun et al. has been published by Jian et al. (2020) for different forest ecosystems across China, although it is less comprehensive. I therefore encourage the authors to clearly state the added value of their dataset compared to previous studies.

Period of the dataset: I acknowledge the effort of the authors to screen 568 publications

and compile all the data, but it would be highly desirable if the dataset could be extended until 2020 so that it would cover a 20-year period (2000-2020), in line with the last two decades of the latest climate reference period (1991-2020). Such a dataset would facilitate the joint analysis of spatio-temporal climate and soil respiration variations (in the context of climate change). Similar to the title for the data repository (Sun et al. 2021), the period of the presented dataset should also be included in the title and abstract of the manuscript. Moreover, it would be helpful to mention the (average) length of the individual time series from the different sites somewhere in the text or to provide a respective figure (e.g. histogram) in the supplements.

Uncertainties: As mentioned in chapter 2.3, most of the soil temperature and respiration data (82 %) were extracted with the WEBPLOTDIGITIZER. This in an interesting approach that provides a workaround to compile scientific data that are not made publicly available in the original studies. However, I was wondering if the authors of these studies have been contacted to request access to the numeric data or was this not feasible due to the number of studies? Were the data from the 568 studies (many of them non-peer-reviewed) checked manually or automatically to identify potential errors or inconsistencies? The given R^2 values of 0.99 for the simple linear regressions (original mean soil respiration data vs. digitised soil respiration data) seem promising, but how does it look like for the monthly data? As a measure for uncertainty, the RMSD or MAE should be provided as well. I am also missing a section in the manuscript that discusses (at least qualitatively) the potential uncertainties originating from the different instruments and experimental setups at the different study sites as well as from the varying time periods of the datasets used for characterising differences between the four climate zones (cold-temperature, temperate, subtropical and tropical). Lastly, are there forest types (e.g. mountain forests) that are potentially under-represented in the dataset (due to a lack of respective studies)? Such a potential bias might affect the numbers provided for the temperature sensitivity of soil respiration and for the annual soil carbon emission originating from forest ecosystems in China. This needs to be to discussed at least briefly.

I would suggest the following modifications for the dataset:

- Use a non-proprietary data format (e.g. CSV file) so that the dataset can be easily read by any software.
- Add a metadata file or readme file that contains all necessary information (e.g. those from Table 1 in the manuscript) so that the dataset can theoretically be used independently of the data paper. Nevertheless, add a reference to the data paper in the metadata/readme file and on the landing page of the repository.
- Create a GeoPackage (.gpkg) or Shapefile (.shp) that contains the metadata (coordinates, elevation, study site name, forest type, etc.) for each study site. Include it in the repository so that it can be easily imported in a GIS by potential users for spatial data visualisation and analysis.
- Add the units (of each column) either in the header or in the metadata/readme file.
- Column "Month": Use the international date format (ISO 8601) or another common date format that can be easier interpreted by a machine (e.g. 2013-07 instead of Jul.,2013). Replace "Month" by "Date".
- No need for column "Period" as the necessary information are already provided in column "Month" ("Date").

- Column “Time”: which time is meant here? No time zone provided. It is unclear to which data the time refers. Split into two columns as well: e.g. “Start” and “End”.
- Remove the tilde in “Age” as this special character is difficult to handle during automatic processing. Alternatively include a column before or after and use a flag (1, 0) to indicate whether the “Age” is measured/precise (e.g. 1) or estimated/approximate (e.g. 0).
- Column “Forest type”: would it be possible to use an integer or acronym for each forest type in the database and provide the full name in the metadata/readme file?
- Columns “Rs”, “T5”, “T10”: Better remove “NA” (leave cells empty) and create another column before or after indicating with a flag whether data are available (e.g. 1) or lacking (e.g. 0). The same for the other columns where NA values exist.
- Column “Annual Rs”: Does this column indeed provides annual averages or rather the mean over the study period? I think it can be deleted as the mean can be easily calculated from the monthly data provided.
- Column “Altitude”: Replace “Altitude” by “Elevation”.
- Although the redundancy may increase the file size of the dataset considerably, I would recommend to copy the metadata (geographic information etc.) into each line (not only in the first row from each site). Otherwise, complications may arise when the dataset is reformatted or analysed. Alternatively, the table could be split into two related datasets. One would include the soil respiration and temperature data and the other one the metadata for each site. An additional ID could be provided for each study site to link the two datasets...

Specific comments

Title: Mention the timeframe of the dataset (“2000-2018” or “2000-2020”) and replace “database” by “dataset”. A database describes a collection of multiple datasets that are generally stored and accessed electronically from a computer system... Title suggestion: “A compiled monthly soil respiration dataset for [various] forest ecosystems across China from 2000 to 2018”

Line 64-73: Maybe just quote the database here and add the URL (with the access date) in the reference list.

Line 75-83: Indicate the period that has been considered. From 2000 until 2018 I think.

Line 85: Do the 568 publications represent 568 study sites or are some data from the same site? Please state the number of considered sites and their geographic and elevational distribution somewhere in the text.

Line 86-91: Why have the other provided variables (e.g. mean annual temperature and precipitation as well as elevation) not been included in the analysis? I am aware that an in-depth analysis is beyond the scope of this data paper, but some additional plots (soil respiration vs. elevation, or soil respiration along selected temperature or precipitation

transects) would emphasise and showcase potential of this dataset.

Line 100: In addition to R^2, provide the RMSD or the MAE as a measure for uncertainty.

Line 105-107: Please clarify whether this sentence describes the procedure in the original studies or how you have modified and analysed the data.

Line 115: I am missing a brief section in the methods about the (statistical) analysis that had been performed to present the data.

Line 119: Provide in addition to the total number of paired measurements (6341 and 2878) also the percentage (in parenthesis) with respect to the total number of considered data.

Line 123-124: Note that R^2 is invalid/inappropriate for non-linear regressions! R^2 cannot differentiate between "good" and "bad" non-linear models. The standard error of the regression could be used instead.

Line 125-132: Uncertainties need to be provided for these values!

Line 134-150: One big problem I see here is that time series spanning different years were used to determine seasonal and geographic differences in soil respiration. I am aware that this is unavoidable when using a compiled dataset, but the uncertainties originating from this issue should be at least discussed qualitatively. Which period do the data cover that were used to compute these values (2000-2018?). Add this information.

Line 149: What do you mean with "winter" in the (sub)tropics? This term does not fit in this context.

Line 154-176: Are the data precise enough to state two decimal places? Confidence in these numbers would be increased if uncertainties were provided.

Line 154-155: Are these mean annual values averaged across all study sites? What's the considered time period?

Line 160: These acronyms were not defined before. Please specify.

Line 165-176: Too many numbers in these paragraphs. I would recommend to provide a comparative figure instead.

Line 177: A general discussion on uncertainties is lacking!

Line 179-196: A brief comparison with values from other regions outside China could be added.

Line 193: See previous comments regarding the use of R^2 and non-linear models.

Line 207-2018: It is difficult to compare any total numbers of Rs (from different sites or studies) if the respective measurement periods, for which these numbers have been computed, are not stated.

Line 214: I would suggest to write the full names and regret from using acronyms if the terms are only used a few times throughout the manuscript as in this case. This enhances the readability.

Line 237: This number conflicts with the total number of Rs data (=10288) stated at the beginning of the manuscript, doesn't it?

Line 258: This sentence is a bit misleading as no in-situ measurement have been performed. Specify that a comprehensive literature review has been conducted to generate the dataset.

Line 251: Any kind of outlook is missing. Does this new compilation for example indicates that there are particular regions or forest ecosystems that are under-represented with respect to soil respiration and temperature studies and deserve more attention?

Figure 1: I would suggest to include more information in this map. A digital elevation model, hillshade, orthophoto or land surface cover classification could be displayed as a background map. Different colours or symbols could be used for the study sites to indicate for example the length of the time series (e.g. 1 year, 5 years, 10 years, 20 years), or the number of available variables at each site (i.e. Rs, T5, T10). The overview map is too

small and has no added value. Better increase or remove it. It would also be helpful to indicate the considered climates and/or forest types.

Figure 2: Note that R^2 is invalid/inappropriate for non-linear regressions (see previous comment). Have monthly or annual data been used for the calculation? From which period do the data originate?

Figure 4: How and for which period were the mean annual fluxes calculated? I assume all annual data from different years and sites associated with one forest type were spatially and temporally averaged. Is this correct?

Figure S1: Does the mean reflects the entire study period at each site? Do the correlations look similar if monthly data (collected vs. digitised) were compared?

Figure S2: I have the impression that there are other non-linear functions that describe the relationship between the soil respiration rate and soil temperature better than the applied ones.