Comment on essd-2020-396

Anonymous Referee #1

Referee comment on "Multi-site, multi-crop measurements in the soil-vegetation-atmosphere continuum: A comprehensive dataset from two climatically contrasting regions in South West Germany for the period 2009–2018" by Tobias K. D. Weber et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-396-RC1, 2021

General Comments

Dear authors,

field observations are the inevitable basis of model investigations as the foundations of calibration and validation. They also form the ground truth for the development of remote sensing applications. Both has the potential for better results with more high quality field data being available. Therefore, for this kind of datasets, originality is not an important criterion. The manuscript under review contributes a large and comprehensive dataset covering relevant observations. The dataset is unique, useful and complete, which make it suitable for publication in ESSD. Most of the methods of data collection described follow community standards. The dataset is well organized in appropriate file formats. The data files are well documented in metadata files and tables. The text is well written with only very few language errors and the list of references is complete.

However, I see some problems with the structure of the text. Resolving these will improve the usability for readers. The current structure of the text makes it more difficult to follow than necessary. I do mainly refer to the differentiation of "field measurements", "field sampling", and "laboratory measurements", which causes data on the same compartments of the system (soil, plants) to be described in more than one place. For a user of the data, it makes it easier to find required information, if all soil data and all vegetation data is described in the same section or in consecutive sections, respectively. Please find additional comments on this in the specific comments.

Furthermore, the text does not cover all of the data provided in the dataset (for1695_data.zip, the file containing the photos, and the supplementary data):

- Soil respiration measurements are in the dataset as is metadata for it. It is also mentioned on the text as chamber measurements or soil respiration. However, the description of the measurement process, devices applied, and so on, is missing.
- gis_data_model: this is a valuable supplement to the data. It deserves to be mentioned in the text.

- Soil profile data: This should be explicitly addressed in the text.
- The soil water figures in the supplementary material may be mentioned in the appropriate position.

Vegetation green fraction, on the contrary, is described in the text but not included in the dataset. This also requires correction. Please see my specific comments on this.

There is only one major deficit of the dataset and the manuscript: There is not sufficient information on data quality. The text lacks descriptions of measures for quality control and assurance as applied during the measurement campaigns and subsequent analyses in the lab or during data analysis. Furthermore, to make the data usable, information on errors and uncertainty need to be presented. This comprises measurement errors of devices used (there is some data included on this aspect), or errors resulting from handling samples in the field and in the lab. For each value in the dataset there has to be a number or a statement on uncertainty. Such a statement may also be an explanation, why no quantitative estimate of uncertainty is given. It has also to be ensured, that flags included in the data files are explained in the text or in the metadata files.

Another information, which is of interest to readers and potential users of the data, is the fraction of missing values. Please consider adding this.

Overall, the manuscript is a good text describing an impressive extensive data set. I am looking forward to the revised version of the manuscript, which will be a valuable contribution even beyond the modellers community.

Specific comments

P2L7: The EC data was aggregated to 30 min.

Fig1: The conceptual model seems not completely coherent. Why are crops shown in in the "land surface" compartment when there is a "crops" compartment? Does the land surface include non-crop vegetation in this conceptual model? And does the crop compartment have its own soil model? Doesn't the crop model simulate plant water uptake, which in turn has to be passed to the soil hydrology model (brown arrow from right to left)? Please make this coherent or comment on it.

P4L15ff: Do not forget the weather data.

Fig2: Please include a (small) map showing where Baden-Württemberg is located in Europe for the international readers.

Sections 2.1 and 1: Consider moving the information on the sites from the introduction to the site description and the information on funding to the introduction

Section 2.1.1 and 2.1.2: The source of the climatic data is missing for KR if Troost and Berger is the source for SA.
Have the sites been used as agricultural fields for a long time? This is interesting for readers from areas where agriculture has begun only some hundred years ago and the soil's slow carbon pools are still influenced by the conversion.
How are soils in SA classified following WRB?

Table3: is this table required? Which additional information is given compared to the data tables? If none, consider removing it. Footnotes 6 to 8 are relevant to users of the data.

Consider adding this in a comment column in the data file.

The information on NA should be in the caption.

Please put the line explaining the asterisk in a separate row in the footnotes. I had to search for it.

What are the numbers in parentheses in the yield per plot column?

Section2: I am missing a section on soil property data. This has been described above but since these are also field measurements (in the same sense as crop measurements are), they should be described here.

Section 2.2: The information on composition of fertilizers in the metadata file is valuable. It should be advertised in the text.

Why is crop yield assigned to management and not to crop/vegetation data?

L12f: Is there an explanation for the discrepancies? What is the uncertainty of the different numbers?

Please state, how the positions of the five plots per site (field) were chosen.

Also refer to the tables in appendix A.

Section 2.3.1: The data files include columns showing flags. The flagging concept needs to be explained here.

Section 2.3.2: The caption of Figure 4 mentions gap-filled data. If gap-filling was done, it needs to be described. Otherwise, the data not included in the dataset should not be shown in the figure.

Section 2.3.3: incomplete

The metadata files mention a calibration step for the soil moisture. This should be mentioned in the text.

A reference to Table 5 is missing.

Section 2.3.4: The heading of the section is inappropriate since it also reports on biomasses and height.

The dataset does not include fresh biomass or vegetation water content. This is valuable information for e.g. radar remote sensing applications. If this data has been measured in an appropriate way, it needs to be added to the dataset. The section also lacks information on how biomass was dried, how much of the harvested biomass was used for LAI determination. Sometimes, blossoms are assigned to generative biomass and only harvestable biomass is reported instead of generative biomass. How was this done in this case?

In the data files on biomass and LAI, I was also missing a column for comments. Sampling of vegetation frequently causes problems with dirty or wet leaves, biomass of weeds, occurrence of pests, and so on. This needs to be reported.

The data files report biomass and LAI per plot. In order to provide the opportunity to use the data with remote sensing data, please provide coordinates of the plots.

The data file on phenology also reports on a replicate. Please explain that in the text. Furthermore, plant height is included in this file but users would look for it in files for biomass or LAI. Please move it there.

Section 2.3.5: Based on the heading of this section, I was looking for GVF data in the dataset. This remained unsuccessful. The dataset only contains photos that can be used to derive GVF. This needs to be clear to readers. Therefore, this section can only be on photos. You may then describe how GVF can be derived from the photos. If I got something wrong and GVF is part of the dataset, the following has to be noted: The hint on the possibility to combine gfv with RapidEye is valuable for readers. However, the results of the study of Bohm et al. (2020) does not belong in this paper. Therefore, remove the sentences from P15L12 on.

This section mentions the biological names of the crops. Because this is not relevant in the context of photos, these names should be removed here. They may be added where the crops are introduced or in the plant metadata file.

Section 2.4: Because this section is on compartments of the system for which other data has already been described above, the part on soil should be added in the soil section. The same applies to information on soil CN measurements, which are addressed here but described in section 2.5.
The information in the plant section is a repetition of section 2.3.4, anyway, and can therefore be removed.

Section 2.4 soil: I cannot find the files for the bare soil plots in for1695_data.zip.
In the second paragraph, information on manual tilling is given. However, it remains unclear, why this had to be done. Were the plots excluded from the management by the farmer? This would mean, that the management data does not apply to the plant data from these plots. Please explain in more detail. Because this information is of relevance for the plant data, mention it in the respective section.

Section 2.5: Analogous to section 2.4, the information in this section should be combined with the other information on the same compartments (soil, plant).

Section 2.5.1: Soil organic C content is assigned the symbol TOC while in the data file it is SOC. From the text, it is not clear to me whether these are different measurrements. The different devices (Vario-EL vs. MACRO Cube) seem to support this. Please rephrase and clarify. I suggest removing TOC from the text if it is not included in the data.
The last line of the section mentions gravimetric estimation of SWC. Because this data is not included in the dataset, remove this information.

Section 2.5.2: Many studies report drying of vegetation samples at temperatures above 100 °C. Please cite a study that shows that 60 °C or only 28 °C is sufficient or report on the comparison of the dry weight before and after determination of residual water content.

All sections: In each section, refer the reader to the data files that contain the respective data.

DATAFILES:

Data: Using NAN as the missing data symbol can cause problems since software will read this as an error instead of a missing number. Missing values should be marked by a different code like e.g. "NA".

Metadata files: There are no column headers for columns 2 and 3. Please add headers for each table or mention in the beginning of each file. If files were formatted using spaces instead of tabs, they were readable independent of the tab-settings of a user's editor.

00_metadata/research_sites_metadata.txt: Missing values are stated to be NAN but missing site names are marked by NN.

00_metadata/management_metadata.txt: L28: What is meant by "Mark the large variability […]"?

00_metadata/soil_metadata.txt: for variable wp add "permanent wilting point" fpr users not familiar with pF values.

00_metadata: names of site-specific soil metadata files include EC, which is misleading. Consider replacing "EC" by "site".

05_management/fertilization/ fertilization.csv: contains three empty row at the end. The last line with data is assigned to site 5, is that right? The related Excel file contains additional sheets.

05_management/cultivation/correction.txt: Correction history only needs to be shown for corrections after publication of the dataset

05_management/soil_management/soil_management.csv: in the metadata, mention the Expert-N is a model

04_weather: double csv in filename

06_plant/biomass/not_publish_calc_biomass.csv: is this file meant to be included in the dataset?

06_plant/lai/ and 06_plant/phenology/ contain a hidden file names ".plot", respectively

06_plant/plant_metadata.txt: the explanation of residual water content may be understood as if this was determined after drying at 60 °C.

Technical comments and details

P1L26: Here and on page 15, the name "rape" is used for Brassica napus. The rest of the text uses "rapeseed"

P5L9: Use the same number of decimals for the coordinates of both sites.

P5L9: The term "research sites" is not consistently used throughout the text

P5L14: […] in 2018, respectively, […]

Table 1, caption: three (?) research sites. Six! Mention, that this is the data from profile_data.csv.

P6L11: Which site is the one ongoing site (EC 1/2/3)?

P7L26: Sentence on coordinates: measurement period should be in parentheses not coordinates since these are the topic of the sentence.

Table 2: Can you mark the crop abbreviations in the legend below the table with the same colors?

P8L5, P8L13: Please correct "(c.f. 0)"

P8L24: Please add area reference to the 0.5 tons (per hectare?).

P8L25: Can you give a reference for the lower yields of group B and C varieties?

P8L26: Please refer the readers to the tables in appendix B in the text where district averages are mentioned.

Table 3: Please explain the meaning of "group" in the caption. Since the table contains fertilization information as the only element of a nutrient balance, the caption is misleading. Instead, yield could be mentioned in the caption. In the legend on crops, what does "E, A, B, C" mean?

P10L3-12: This belongs to section 2.3.2. Information is partly repeated there.

Table 4: There is no need to cite here without referring to additional relevant information on instrumentation in that paper.

Figure 3: "top" missing in caption. What does "SJ" refer to? Description of the data shown should be moved to the text.

P11L12f: The sign convention should also be mentioned in the metadata file for weather.

P11L13f: This sentence seems to be related to Figure 3 but no reference to the figure is given. In addition, there is something missing before the comma.

P11L16: Figure 5 rather belongs to section 2.3.3

P12L29: There is more in the figures than just data coverage.

P12L30: "Supplementary" instead of "supporting"

P14L9: Hukseflux

P14L13: There is no data on soil temperature or moisture in Table 3.

P14L24: Please explain what "five extra plants" means. Why are they "extra"?

Figure 6: The figure is not consistent with the data in for1695_data.zip. The same applies to the readme file in for1695_data_GVF.zip

P17L14: This applies to comparisons with the aim to find effects of the location only.

Tables 7ff: Consider unifying the beginning of the captions with or without "determined".

Terms: Please unify "matric potential" and "matrix potential". Both are used in the text or metadata files.

Tables in appendix B: Mention source of the data in the caption.

Please check the tables in section 3 and the metadata files for consistency.

References for citations in metadata files (like Priesack, 2006 or Foken, 2006, van Genuchten, 1980, Ad-hoc-AG Boden, 2005) need to be given either in the metadata files or in the list of references in the text, or, maybe preferably as a service to the users, in both places.