*Interactive comment on* "Long-term Field Experiments in Germany: Classification and spatial Representation" *by* Meike Grosse et al.

**Anonymous Referee #2**

Received and published: 3 July 2020

Grosse et al. compiled an overview over historic and still running long-term experiments in Germany. I applaud this effort because it will allow better use of these expensive and precious experiments.

However, their assessment could be more critical and identify major deficits of the present LTFEs. As far as I can see, all are situated in flat areas (a data evaluation in this respect would be nice and not too difficult to do). This means that they exclude major lateral processes (interflow, surface runoff) and differ largely from typical agricultural fields, in contrast to the conclusions from the climatic water balance during the growing season (CWBg) and the Müncheberger Soil Quality Rating (MSQR) that seem to indicate representativeness. This deficit may be especially pronounced for grassland experiments because grassland either occupies lowland areas that are too wet

for arable use or areas that are too steep. For grassland experiment, which in fact are meadow experiments (grazings seems to be missing; also a major deficit). Such critical assessment would be extremely helpful to guide the installation of future LTFEs and to show the limitations in the conclusions that can be drawn from the existing LTFEs.

Were lysimeter experiments included, which would allow assessing at least vertical water fluxes? Do long-term experiments with lysimeter exist at all in Germany? Again, a critical assessment would be helpful. Were experiments included that allow quantification of lateral processes (runoff, soil loss)? I could imagine that the measurements in Trier (Stehling and Schmidt 2017) or those by Jung and Brechtel (1980) qualify for LTFE. If they don't qualify, this would again illustrate a major deficit of present LTFEs.

In the discussion I missed a wider view. Do similar compilations also exist in other countries? Are the German LTFE experiments similar to what was done and is done in other countries?

Furthermore, the authors give the impression that they still focus on the old questions of LTFEs (mainly yield) that became boring. I had this impression for two reasons. First, little examples are given how LTFEs can be used in fascinating modern research on urgent questions. Second, using LTFEs in modern research applying new techniques requires access to the experiments. Hence it makes a big difference whether an experiment is still ongoing or not. However, this information is given nowhere. Second, it often requires archived samples (as an example what can be done with modern techniques and archived samples, Köhler et al. 2012 comes to my mind but there are certainly more examples). This information, whether archived samples are available, should be included. Generally, I missed information about which data could be obtained from the LTFEs.

Most of my other remarks are mainly editorial issues. The weakest part in this respect is the table in the Appendix, which is most important because it resolves the LTFEs and thus allows access (see below).

Details (numbers refer to lines):

12: add "during the growing season"; I would even change the abbreviation to CWBg because usually an entire year is considered in a CWB. I was very surprised when suddenly somewhere in the manuscript the information 'growing season' popped up

13: Müncheberger Soil Quality Rating seems to be a combination of German and English. Shouldn't it be 'Müncheberg'?

35: I welcome this definition of the control that is certainly better than the often used but wrong assignment of the strongest and most unrealistic intervention as control, namely the long-term nutrient removal. However, I did not find this definition to be used later in the manuscript.

46: Bai et al.

116: Not clear how PET was derived. Was it taken from DWD? Is it Haude?

126: This is strange. Later only 6 classes of the MSQR are used, not 102. I wonder whether different properties like soil structure, wetness, relief, contaminations can be combined in one indicator of six classes. This may be possible for one specific target like yield but will fail for most other targets or require other classes. Is a better resolution than these six classes possible?

128: I guess this should read 'available water capacity'

130: What is unsuitable? This always requires the definition of a target.

139: This leads to the question: Were lysimeter experiments included? If not, why not?

155: The title does not have this restriction; also the Abstract does not. I wonder why it suddenly pops up in the results. I also wonder how this is defined (what is bioeconomy?) and whether these experiments really aim at sustainable soil use. They exclude many things that make soil use unsustainable (erosion, compaction) and hence are unsuitable to test sustainability (in this general sense). I also wonder even more why

the criterion sustainability excludes some grassland experiments. This is contrary to what I would expect.

160: Establishment was in the past. Hence past tense would be appropriate. The question of correct tense is rather difficult to answer given that 30% of the experiments have come to an end already and others will come to an end in the future, I wonder whether the mostly used present tense is justified.

171-172: One sentence is usually not a paragraph. Furthermore, temporal aspects were treated in the first paragraph of the results. I suggest moving this sentence.

173: sentences usually do not start with a number; this also applies in other cases (e.g. L. 181, 184).

178 : Move opening parenthesis

208-209: This should be moved to the M & M section; this is the first time that growing period is mentioned although CWB appeared already several times. Furthermore, it would be good to explain the rationale behind this decision than let the reader speculate

266-269: I would reverse the argument. In my view the critique by Franko is well justified and shows that 6 classes of the MSQR are insufficient. I do not suggest to include an assessment of the complexity of soil parameters but it is also not justified do say that the LTFEs are representative regarding soils just because they match the rather coarse and restricted (to yield) MSQR criterion. References: The format varies among references. Please homogenize

Fig. 2: The pie charts are an attempt to illustrate the manuscript. However, they do a poor job. They require a legend, which is difficult to read (because font size is smaller than that of ordinary text) and contain information that is better suited for a table or even could be given as plain test. For Fig. 2 a, a density graph would be more appropriate

Fig. 3: A graph usually has not a title but a caption. The colors are impossible to distinguish Are they necessary? Can they be simplified? Wouldn't the year when an

LTFE was closed be equally interesting?

Table 1: It is not clear whether 'organic fertilization' also includes straw and compost (there is not an equivalent 'Mineral fertilization'). Furthermore, why are green manure, compost and sludge mentioned, but not the main type of organic manure. This classification appears inconsistent. It surprises me that only two of the grassland experiments have organic fertilizer although grassland use unavoidably produces manure. Have all except for two experiments used an unrealistic design that does not allow application of the results to typical situations? Better call 'plant protection' 'crop protection'

Fig. 4: same remark as Fig. 2

Table 2 + 3: 'vegetation period' should not be in the column head but in the caption. Also the lines separating groups of variables are not consistent (why are CWB class and range separated by a line? Isn't the unit for CWB mm/yr?

Fig. 5: Here four classes of LTFE are sufficient. Why does Fig. 3 require eight classes (that cannot be read anyhow)? LTFE should not be repeated five times in the legend. It is not necessary at all. CWB is in mm/yr

Fig. 6: Delete LTFE

Fig. 7: column widths could be much smaller while larger row heights would allow a larger font size. Presently the numbers hardly can be read. It is not necessary repeating 'MSQR class' six times. Better use a larger font size. The colors of the legend should agree with the colors in the graph.

Table A 1: This is likely the most important table because it allows access to the LT-FEs. However, it is rather inconsistent and difficult to read. E.g., the IDs cannot be read; some institutions got abbreviations (why?) others not; some places are mentioned, others not (why?). Mentioning the main institution may be fine in hierarchical organizations but this is clearly insufficient for big universities. Whom should one ask there? I suggest replacing the information in column 3 by a number and the place and

resolving the number below the table by reporting the full addresses. This would also create room for the other columns. Furthermore, I see no reason why umlauts are replaced. This is poor technology of the past century and again a waste of space.

Jung, L. & Brechtel, R. (1980): Messungen von Oberflächenabfluß und Bodenabtrag auf verschiedenen Böden der Bundesrepublik Deutschland. – Schriftenreihe des Deutschen Verbandes für Wasserwirtschaft und Kulturbau (DVWK), 48; Parey, Hamburg

Köhler I, Macdonald A, Schnyder H (2012) Nutrient supply enhanced the increase in intrinsic water-use efficiency of a temperate seminatural grassland in the last century. Global Change Biology 18, 3367–3376, doi: 10.1111/j.1365-2486.2012.02781.x

Stehling, E. & Schmidt, R.-G. (2000/2017): Das Datenarchiv der Forschungsstelle Bodenerosion in Mertesdorf (Ruwertal). Eine Dokumentation über 25 Messjahre (1974 - 1999); Informationszusammenstellung zum Gebrauch der Daten-CD.