Response to reviewers
First review round
September/October 2022

Title: Quality Assessment of Meta-Analyses on Soil Organic Carbon

Authors: Julia Fohrafellner, Sophie Zechmeister-Boltenstern, Rajasekaran Murugan, Elena Valkama

---

John Koestel, 09 Sep 2022

I am thanking the authors of egusphere-2022-488 for this exceptionally well-developed manuscript, in which they investigate the quality of available peer-reviewed meta-analyses on soil organic carbon. I only have some one major point that I want the authors to address. In other words, I recommend a publication after minor revisions.

There is mismatch between the literal meaning of the term “meta-analysis” and the meaning it has for the authors of this manuscript. The literal meaning of “meta-analysis” is something like “transcending analysis” or “analysis of “several” analyses”. As such, the term does not imply whether this transcending analysis is qualitative or quantitative or which statistical methods need to be used for a meta-analysis. Some examples of the definition of meta-analysis in dictionaries are

“a research method that combines the results of several related studies to produce better results” (Cambridge Dictionary)

“Meta-analysis is a statistical process that combines the data of multiple studies to find common results and to identify overall trends.” (Dictionary.com)

“a quantitative statistical analysis of several separate but similar experiments or studies in order to test the pooled data for statistical significance” (Merriam-Webster)

Also, the definition given by the authors (L36-41) does not specify what kind of statistical methods need to be used in a “proper” meta-analysis.

I very much agree with the authors that there are more and less suitable methods. However, I am unconvinced that the three cut-off criteria (effect size, standard deviation, studies weighted by 1/variance) defined by the authors are suited to distinguish between correctly and wrongly executed meta-analyses. In some fields, the available data in the literature does not allow for investigating effect sizes, since the studies were not carried out using a design that allows classical statistical testing. An example for such a field are soil physics. Here, respective measurements are so time-consuming and the number of “moderators” so large that studies have been carried out with the goal of process understanding. Still, statistical analyses (e.g. correlation analyses or machine learning approaches) can be applied and still deliver useful insight. Likewise, there are cases were the correct standard deviation of repeated measurements are not given by the original publications (e.g. in cases were the data is log-normally distributed). Using the number of replicates (or a thereof derived measure) as a weight instead of 1/variance maybe the only possible way to weight the data

in a more reasonable fashion. Please also consider that the variance may not only reflect measurement error, but also express small-scale heterogeneity in soil properties. In the case of organic carbon measurements, pooled satellite samples can be measured and the problem be circumvented. In the case of tension-disk infiltrometer measurements, each point in the field needs to be measured individually. Following a randomized block design would often be too expensive. Therefore, replicated measurements in publications are often only pseudo replicates. Provided there are sufficient (pseudo)replicate measurements, the geometric mean hydraulic conductivity at a plot with large small scale heterogeneity will reflect the hydraulic conductivity similarly well as the one in a field with small small-scale heterogeneity. Only its variance will be much higher. Using 1/variance as a weight will then up-weight locations with small soil heterogeneity not the once with the better measurement precision.

I am not per se opposed to the idea that a proper meta-analysis must include effect sizes calculated from measurements weighted by 1/variance. If the majority of the meta-analysis community defines "meta-analysis" in this fashion, this may be the way to go. It is however very probable that the authors of the here reviewed "meta-analyses" had not been aware of this very strict and specific definition. I therefore urge the authors to better explain why a meta-analysis must include effect sizes calculated as outlined above and suggest than a term for studies that also analyze the results of several source publications using statistical methods, albeit not the ones required by an orthodox meta-analysis. Alternatively, I suggest the term "meta-analysis" as a broader term and the term "orthodox meta-analysis" for conducting a meta-analysis in the sense of Koricheva and Gurevitch (2013).

A minor thing: the plural of meta-analysis should be meta-analyses. At some places in the manuscript it is correctly spelled, at others it is not.

We want to thank Dr. Koestel for reviewing our manuscript and for his very positive but also critical feedback.

We know that there is a debate on what a "meta-analysis" is, and we will include this in the introduction to give a clear picture on how and why we position ourselves as we do. It is true that the definition of "meta-analysis", which was coined by Glass in the year 1976, is more general than we portrait here in our manuscript. In the past, and still today (at least in soil and agricultural sciences), this term is used to describe synthesis efforts in a broader way, just as the definitions mentioned above. But this is part of the reason why, still, "orthodox" or as we like to call them, "true" meta-analyses of low quality are being published. Scientists in our field often do not know what constitutes a true meta-analysis; that a summary effect sizes needs to be calculated, based on independent studies which are weighted. This method follows strict rules, which have been defined by experts in biostatistics (Borenstein et al. 2009; Cooper et al. 2019; Koricheva, Gurevitch and Mengersen, 2013). In other fields (medicine and social sciences), where meta-analyses and systematic reviews have been first developed and applied, both methodologies are standardized and clearly defined. Every other form of quantitative synthesis which does not follow these rules, cannot use these terms to identify themselves. We hope soil and agricultural sciences will follow these research areas and start to clearly distinguish between the forms of (quantitative) reviews, instead of using terms regardless of methodology applied. This is, in our opinion, a critical aspect when it comes to solving the quality issues we currently see. Without clear definition,

distinction and criteria, it will not be possible to overcome this problem. We completely agree with the comment that when effect size calculation and moderator analysis are not feasible, correlation analyses or machine learning approaches can be useful. Nevertheless, we urge to not call these analyses "meta-analyses", but e.g. quantitative analyses, correlation analyses, analyses through machine learning, etc. Similar claims have been done in ecological and conservation sciences. Cote & Reynolds (2012) said that "the term 'meta-analysis' has been [incorrectly] applied broadly to any analysis that includes data from multiple sites, species, and/or sources, regardless of the nature of the response variables considered or the analytical procedure used". Likewise, Koricheva and Gurevitch (2014) write that "Another common misconception seems to be that the term 'meta-analysis' can be used whenever data from several studies have been extracted and analysed in some way. Metaanalysis has its own well-established methodology, which is described in numerous textbooks [Borenstein et al. 2009; Cooper, Hedges & Valentine 2009; Koricheva, Gurevitch & Mengersen 2013a]". Vetter, Rücker & Storch (2013) write: "It could be argued, that the term 'meta-analysis' has developed a different tradition of application and perception in ecology and conservation biology than in the medical sciences. Instead of the rigorous technique of medical sciences, meta-analysis in ecology and conservation biology may rather refer to a more general type of analysis at a meta-level". They further say "Seemingly, researchers connect the term meta-analysis with the idea of some quantitative statistical calculations combining independent studies from the literature; an idea that is not wrong in itself, but too vague and not sufficient to be qualified as a meta-analysis. Repeatedly, multiple regression studies from the journal Ecological Economics [Brander et al. 2007, Richardson and Loomis 2009, Barrio and Loureiro 2010] or correlational studies were termed meta-analysis [Hartley and Hunter 1998, Benayas et al. 2009, Creel and Rotella 2010]".

Gurevitch et al. (2018) describe meta-analysis in steps as follows:

1. "In a meta-analysis, one or more outcomes in the form of effect sizes are extracted from each study. Effect sizes are designed to put the outcomes of the different studies being combined on the same scale, using a suite of metrics (e.g., logarithmic ('log') response ratios)".
2. "The effect sizes are then entered into a statistical model with the goal of assessing overall effects and heterogeneity in outcomes. These models are based on an assumption of either a common effect ('fixed effect') or random effects".
3. "In carrying out a meta-analysis, the central tendency (the mean) and its confidence limits are evaluated, as well as the heterogeneity in the effect across studies".
4. "To identify the magnitude and sources of variation in effect size among studies, earlier studies relied on simple heterogeneity tests, whereas more recent work often uses meta-regressions".
5. "Heterogeneity tests and meta-regressions both use weighting based on the precision of the estimate of the effect: larger studies with higher precision are weighted more heavily than smaller and/or more variable studies".

Regarding weighting, Cooper et al. (2019) describe it by 1/variance (Eq 12.2, p.247). Hedges et al. (1999) write: "... the effect estimates from different experiments will typically differ in precision (standard error). Therefore a weighting of the individual study estimates giving greater weight to experiments whose estimates have greater statistical precision (smaller

standard error) will increase the precision of the combined estimate" (see Eq. 5). Their views align with the ones of Borenstein et al. (2009) and Koricheva et al. (2013).

References:

1. Borenstein, M., Hedges, L. V., Higgins, J., and Rothstein, H.: Introduction to meta-analysis, Wiley, 421 pp., 2009.
2. Côte, I.M. & Reynolds, J.D. (2012) Meta-analysis at the intersection of evolutionary ecology and conservation. Evolutionary Ecology, 26, 1237–1252.Cooper, H., Hedges, L. V., and Valentine, J. C.: Research Synthesis as a Scientific Process, in: The Handbook of Research Synthesis and Meta-Analysis, edited by: Cooper, H., Hedges, L. V., and Valentine, J. C., Russell Sage Foundation, New York, 4–15, 2019b.
3. Gurevitch, J., Koricheva, J., Nakagawa, S. *et al.* Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018). https://doi.org/10.1038/nature25753
4. Hedges, L. V., Gurevitch, J., and Curtis, P. S.: The Meta-Analysis of Response Ratios in Experimental Ecology, Ecology, 80, 1150, https://doi.org/10.2307/177062, 1999.
5. Koricheva, J., Gurevitch, J., and Mengersen, K. (Eds.): Handbook of Meta-analysis in Ecology and Evolution, Princeton University Press, 520 pp., 2013.
6. Koricheva, J. and Gurevitch, J.: Uses and misuses of meta-analysis in plant ecology, J. Ecol., 102, 828–844, https://doi.org/10.1111/1365-2745.12224, 2014.
7. Vetter, D., R€ucker, G. & Storch, I. (2013) Meta-analysis: a need for well-defined usage in ecology and conservation biology. Ecosphere, 4, 74.

Bas van Wesemael, 13 Oct 2022

The paper is well-written. It uses established techniques for meta-analysis developed in other domains an verifies to what extent these are respected in agricultural and soil science.

We want to thank the reviewer for his time and the critical reading of our manuscript. The suggested improvements were highly appreciated and all adopted into the manuscript. Below, we respond in more detail for comments including questions to the author team.

Lines 152-153 please break up this complex sentence and specify more clearly what your example 'such as the IPCC report' refers to.

Line 169 Please check the consistency of spelling: 'criterion' or 'criterium'?

Lines 191 I assume that all inclusion and exclusion criteria apply for the search strategy, even though 4-6 are not discussed. It would be helpful to stress that all criteria haven been checked.

Line 205 'respectively' can easily be avoided: e.g. '344 in Web of Science and 208 in Scopus'. This facilitates the reading.

Line 209 Please refer to Table 2 for these criteria.

Line 216 I would suggest to use 'evaluated' instead of 'analyzed' in order not to use 'analy…' too much in one sentence.

Line 254 It would be helpful to remind the reader which scores.

Line 267 It is not clear that 'group' refers to the set of criteria in Table 1. As the table is quite large and on multiple pages, these groups are not evident.

Line 279 Please refer to table 1 to clearly link the groups to the text.

Line 295 Please delete 'studied'

We changed the manuscript according to your comments (lines 152 to 295).

Line 296-297 Do you have any explanation why the meta-analysis that responded to the strict 'cut-off criteria' showed a much poorer performance than the other ones? Please so remind the reader that the cut-off criteria are criteria 6-8 in Table 1. I had to look it up again.

This figure shows that when considering only meta-analyses that fulfill the cut-off criteria (=true meta-analyses; striped bars), a lower % of studies complies with the quality criteria 9-17. Simply because the number of true meta-analyses is only 4, while "pseudo" meta-analyses is 27. Here we want to display that when including the pseudo-meta-analyses, the fulfillment of quality (%) might look better than when actually only considering true meta-analyses. As another reviewer also mentioned that our display is unclear, we will change the figures and display the results as stacked bars, including a new legend. Misunderstandings should be avoided that way.

Line 368 Did you check if there were any of the 31 papers identified earlier among these 16 papers in the IPCC report?

No, this chapter of the report did not cite meta-analyses that were part of the 31 we analyzed.

Line 490 Please specify more clearly what 'all four reviews' refers to. Maybe refer to table 5. If you mean the four papers cited in lines 492-493, it should be five reviews (including your own).

Yes, correct, we changed it accordingly.

Tommaso Tadiello, 25 Aug 2022

Congratulations to the authors! This is a great work! It will help the scientific community to evaluate the meta-analysis demonstrating their reliability.

Here just a few comments and suggestions:

- Cut-off criteria: great idea! Especially because they give the opportunity to evaluate meta-analysis quickly. A couple of suggestions: (1) I was wondering about the importance of including the “control and treat described” and “independence of effect sizes” in the cut-off criteria group. The description of the control and the treatment is essential for the comprehension of the single meta-analysis. The independence of the observation is really important as well and, in my experience, is frequently not met. (2) Table 1: you probably ordered the criteria based on the usual meta-analysis construction (i.e., starting from the literature search). I agree with that. But, at the same time, the 3 cut-off criteria are not enough in evidence to me. Solutions: (I) put the cut-off criteria at the beginning of the table and then list all the others; (II) leave the table as it is but somehow highlights the 3 cut-off criteria (III) create a separate table just for them. It would be as in Table S6.

Table 1, criteria 4: there's one more space at the beginning;

- just a comment: figure 5 is wow! Such a poor quality so far!!

the fig share link doesn’t work, please review it.

We highly thank the reviewer for the positive comments and feedback.
Regarding the cut-off criteria, we appreciate suggestions on how to highlight their importance. The reason we chose “only” these three criteria as most crucial, is due to the fact that without fulfilling them, a synthesis cannot be called “meta-analysis”. Of course, other criteria, as “control and treatment described” and “independence of effect sizes” are important and if not kept, lead to lower quality, but a meta-analysis is still recognized as such. We also see that even in recent meta-analyses, these criteria are often not met (e.g., Li et al (2020), Mathew et al. (2020)). The suggestion to highlight the cut-off criteria visually is a good idea, which we will take up into Table 1.

References:

Li, Y., Li, Z., Chang, S.X., Cui, S., Jagadamma, S., Zhang, Q., Cai, Y., 2020. Residue retention promotes soil carbon accumulation in minimum tillage systems: Implications for conservation agriculture. Sci. Total Environ. 740, 140147. https://doi.org/10.1016/j.scitotenv.2020.140147

Mathew, I., Shimelis, H., Mutema, M., Minasny, B., Chaplot, V., 2020. Crops for increasing soil organic carbon stocks - A global meta analysis. Geoderma 367. https://doi.org/10.1016/j.geoderma.2020.114230

Damien Beillouin, 29 Aug 2022

General comments:

This paper investigates the quality of meta-analyses published on the effect of agricultural practices on soil carbon in Europe. This is an important and timely topic as the number of meta-analyses is increasing strongly. It is therefore timely to reach out this research community to raise awareness of these issues.

The paper is overall well structured, and well referenced, with an introduction clearly defining the terms and objectives. For the material and method, the authors have adapted quality analysis grids from meta-analyses used in others scientific fields. The authors repeatedly emphasise the specificities of soil science work that would require specific meta-analysis methods or analysis grids. I was not convinced by these arguments, at least, as currently formulated. I also have some questions about some of the 17 criteria used to judge the quality of meta-analyses. The authors identified 31 published meta-analyses dealing with agricultural practices on soil carbon in Europe (on cropland). Recent work has identified more than 100 meta-analyses on these issues worldwide. Given that most of the meta-analyses are global (across continents), this low number identified by the authors raises questions for me.

Dear Dr. Beillouin,

We greatly appreciate you taking the time to read our manuscript and respond in such detail. Your expert opinion and input are of great value. The author team is glad you agree that this topic is important. We took the liberty to respond to the comments you raised in your introductory statement within the detailed comments you wrote per line.

Detailed comments:

Introduction :

L 48: “Particularly, the use of meta-analysis as a tool to investigate the effects of agricultural management practices on relevant response variables, such as yield or soil physical or chemical parameters, is becoming increasingly prominent (Valkama et al., 2019, 2015).”

- Are the two references particular example of meta-analyses on this subject? If so, how do you choose these references? (the first meta-analyses in the academic field?).
- Or do these references analyze the interest and use of meta-analyses in agricultural field?

We changed the references to recent meta-analyses studying these parameters.

L 52: ”Because of their close relationship, many methodological applications of meta-analyses in ecology are also transferable to the field of agriculture and soil science”

- What do you call “applications of meta-analyses”. Does this refer to the method itself? Or the use of the results of meta-analyses?
- Doesn't this contradict your arguments about the specificity of soil carbon meta-analyses?

Here we want to emphasize that many reasons for doing meta-analyses in ecological research are transferable to agriculture, as e.g. combining results of field experiments, across several sites or assessing the impacts of environmental drivers (follow up sentences). The idea is to point out similarities and then explain why there are specific challenges for soil and agricultural sciences.

L 54. “provide clarification by synthesizing conflicting evidence from primary studies. “

- Mention that this clarification is provided by the increase statistical power/increase precision of the mean estimates of the effect ?

We will describe this in more detail to improve understandability. We want to express the issue that experimental studies, looking at a specific treatment compared to a control, sometimes report contradictory outcomes compared to similar studies. By including all studies and calculating a summary treatment effect, meta-analysis allows us to combine all available knowledge, regardless of their outcomes, and calculate one number, which tells us about the overall estimated effect, thereby overcoming conflicting evidence.

L 56. “Nevertheless, research on agriculture and soil encounters issues, which are often specific to these fields. Firstly, changes in soil, like soil organic carbon (SOC), are often slower and more difficult to detect due to smaller sample size and the magnitude of changes (Mäkipää et al 2008) than other physiological and biogeochemical changes; e.g., changes within plant tissue. Moreover, changes in SOC due to management practices have different responses depending on soil depths that need to be taken into account when summarizing results across studies.

Doesn't this need of long term experiment also apply to fields in ecology (e.g. effect of disturbances on the composition/recovery of populations,...) and medicine (e.g. effect of environmental conditions/substances on the prevalence of some cancers,...), or other scientific fields?

This is true, we changed the text in the manuscript.

L 61. “Therefore, it is crucial to define not only the treatment but also the control of the experiments precisely to allow computation of heterogeneity”

- I totally agree, but I don't think this difficulty is specific to agronomy studies. In fact, all effect-sizes are metrics quantifying the relationship between two entities (generally the control and the treatment).

From our experience, the definition of the control is quite challenging and not well understood by many soil and agricultural researchers that apply the meta-analytical method. This is due to the complexity of management practices and their interaction across studies,

and each management practice is affecting SOC. When the control is not defined clearly, this can lead to the inclusion of studies with controls differing from each other (different fertilization, different tillage regimes, grassland/cropland/forest/horticultural experiments combined in one analysis, field/greenhouse/lab experiments, etc.). This is a mix of "apples and oranges", which can cause unreliable results, as it is not possible to distinguish whether the effects on the response variable are caused by the management practice studied (and moderator effects of interest) or are influenced by the differences between controls.

L 64. "A good example is bulk density, which can be measured in a field experiment or estimated using pedotransfer functions in order to compute SOC stocks from concentrations"

- An other example could be the soil depth, sometimes very different across studies, and not always precised.

Bulk density is an observation data and soil depth is not. But soil depth is needed to derive bulk density! They are closely related and add complexity in carbon storage calculations.

A uniform depth, and therefore bulk density measurement can be applied to all sites. But application of pedotransfer functions from one place to calculate BD cannot be applied to another site.

L 87. "However, they are formulated rather generally"

- Could you precise what generally formulated means? Philibert and Beillouin papers presented 8 and 20 quality criteria, respectively (defined in table 1and Figure 6 of their article respectively).

We believe that the criteria presented in the study of Philibert (2012) are not extensive enough for the quality assessment of soil meta-analyses, as they e.g., do not mention the issue of study non-independence, which is very common in our research area, or provide clarification on what makes a meta-analysis repeatable or transparent regarding in- and exclusion criteria or database. The publication by Beillouin et al. (2019) presents a well-structured and extended criteria-set. Nevertheless, we think only experienced researchers can understand the criteria without further investigating how they apply in a specific case (e.g., how is heterogeneity analyzed correctly; what makes a study independent; weighting by accuracy should be done only by 1/variance; the dataset should include all data necessary to reproduce study). In our opinion, inexperienced researchers need a more extended criteria-set, which includes all relevant components of an agricultural or soil related meta-analysis. Additionally, more detailed descriptions and direct quotes from literature are valuable to assist stakeholders in understanding and using the criteria correctly. The criteria-set in our manuscript, especially the extended version in the supplementary material, provides such sections. In the discussion chapter we go even more into detail why we argue for these criteria and what mistakes are made commonly. We will also add relatable examples for soil and agricultural researchers there. By doing so, we want to enable all researchers of our field, even with no or little prior knowledge on meta-analysis, to understand key elements and recognize them, which we believe is not possible with the criteria-sets available at the moment.

L 93. “but do mainly focus on systematic reviews and maps and contain elements not necessary in meta-analysis (e.g. registration, gathering a maximum of available relevant literature or performing critical appraisal)”

- Could you justify that registration, gathering a maximum of available relevant literature or performing critical appraisal are not required for meta-analyses concerning soil organic carbon.

Systematic reviews are required to follow the PRISMA checklist (https://www.prisma-statement.org/documents/PRISMA_2020_checklist.pdf). Regarding registration, CEE expects authors of systematic reviews to register the title and protocol at the PROCEED homepage (https://www.proceedevidence.info/). Similar procedures are required in medical and social sciences (Cochrane https://training.cochrane.org/, PROSPERO https://www.crd.york.ac.uk/prospero/). It is further obligatory to critically appraise selected studies when conducting a systematic review (CEE, 2018). Both are not obligatory for meta-analyses, as it is the statistical procedure itself (Borenstein et al. 2021) and therefore part of the systematic review. But it also can be conducted independently, of course outcomes are only meaningful if the studies have been collected in a systematic way, which is not necessarily done according to the rules of a systematic review. It is also possible to e.g. synthesize “[…] data from a selected group of studies, such as those conducted by a pharmaceutical company to assess the efficacy of a new drug” (Borenstein et al. 2021, p. xxviii). Therefore, it is not always the aim of a meta-analysis to gather a maximum number of studies (or even all relevant data available), but to quantitatively analyze data of e.g., specific trails, only scientific or only gray literature of certain databases.

- For example, the same research question can be addressed in parallel by several research teams, duplicating the necessary research efforts and investments on the same topic (could also occur in agronomic/soil science academic field)-> registration could potentially allow avoiding these problems.

- + in the discussion you mention “The publication of protocols prior to a meta-analysis would benefit the method by allowing constructive criticism and suggestions for improvement by the scientific community (Moher et al., 2015; Brandt et al., 2013).” Is this sentence coherent with the one L 93?

Although the registration or publication of protocols is not obligatory for soil and agricultural meta-analysis, we nevertheless encourage the publication, as we describe in the discussion.

- gathering a maximum of available relevant literature is generally recommended (e.g. by the search of several bibliographic databases). Incomplete base of scientific paper in meta-analyses could lead to wrong results.

Looking at the criteria #1 in Table 1, you give a higher score to meta-analyses using >4 databases. This contradict thus your above-mentioned sentence?

This depends on the data the meta-analysis is aiming to analyze. If the aim is to find the summary treatment effect of a large trial including many sites conducted by an institution, there will only be one database. As the goal is not to provide estimates of the treatment

effect, which are supposed to be communicated as regionally, nationally or globally relevant, but as findings of this specific trial, it is fine to "only" have one database. Contrary, when aiming to find effect estimates that are relevant on a bigger scale and supposed to inform various stakeholders or policy makers, the underlying data should be broadly searched. This is the case of most meta-analyses performed in our field. Therefore, we suggest to score database availability according to criteria #1.

- The quality of the primary research is very variable. Weighting the evidence according to the (estimated) quality of each paper could avoid to give a large importance to the papers with the lowest quality.

Firstly, we want to note that under critical appraisal we understand "the stage at which the individual studies included in the review are assessed for their reliability for answering the Systematic Review question" (CEE, 2018, section 7 Critical appraisal and study validity). This does not include the weighting of studies, which is crucial in meta-analysis and part of our cut-off criteria.

Secondly, we would like to cite Borenstein et al. (2021, p. 416): "Rather than thinking of meta-analysis as a process of garbage in, garbage out we can think of it as a process of waste management. A systematic review or metaanalysis will always have a set of inclusion criteria and these should include criteria based on the quality of the study. For trials, we may decide to limit the studies to those that use random assignment, or a placebo control. For observational studies we may decide to limit the studies to those where confounders were adequately addressed in the design or analysis. And so on. In fact, it is common in a systematic review to start with a large pool of studies and end with a much smaller set of studies after all inclusion/exclusion criteria are applied."

References:

- Collaboration for Environmental Evidence. 2018. Guidelines and Standards for Evidence synthesis in Environmental Management. Version 5.0 (AS Pullin, GK Frampton, B Livoreil & G Petrokofsky, Eds) www.environmentalevidence.org/information-for-authors.
- Borenstein, M. et al. 2021. Introduction to Meta-analysis. Second Edition. Wiley, Oxford, UK.

L 137. "Moreover, the interest in SOC sequestration and subsequent increase in related publications raises the question whether there are meta-analyses synthesising this knowledge"

è A (partial) answer could be found in "A global overview of studies about land management, land-use change, and climate change effects on soil organic carbon" Published in 2021. The authors search exhaustively the literature to identify meta-analyses on soil organic carbon. They identify 192 meta-analyses, and then characterized the interventions, outcomes, temporal dynamics of publication and spatial distribution of the primary studies synthetized in these meta-analyses. They also analyzed Also some quality criteria of the analyses (based on indicators found in Phillibert et al., and Beillouin et al.). See for example Fig 5 of their publication and paragraph 3.3.

The study by Beillouin et al. 2021 is a great output which was not known to the authors until this point and will be taken up as a reference in the manuscript. It will not only be highly useful for the introduction but also discussion. Regarding the question of the entry statement "The authors identified 31 published meta-analyses dealing with agricultural practices on soil carbon in Europe (on cropland). Recent work has identified more than 100 meta-analyses on these issues worldwide. Given that most of the meta-analyses are global (across continents), this low number identified by the authors raises questions for me", we would like to respond here. After reading the mentioned article (Beillouin et al., 2021) and looking at the database, it is now clear why this study identified a greatly higher number of meta-analyses than this review. Beillouin et al. (2021) also included all globally available meta-analyses looking at "land-use management" effects on 1) SOC or 2) other response variables as e.g., yield or CO_2, and SOC as a covariate, and 3) includes grassland, wetlands, forests and cropland, whereas this review has a narrower scope. It is described in the response below to comment of L 148.

L 148. "This study aims to quantitatively analyze 31 meta-analyses, "

- Does these 31 meta-analyses represent a subset of all meta-analyses published on Soil organic carbon focusing on cropland, and European region?. How did the authors define that a meta-analysis focused on Europe when most published meta-analyses use global data? did they exclude all meta-analyses that included data outside Europe? Or on the contrary, did they include all meta-analyses with at least one primary study in Europe?

The latter is true. Within our review, we only included meta-analyses studying the effects of various agricultural land management practices on SOC in mineral soil cropland (SOC needed to be the response variable and not another covariate which was analyzed additionally). Among the studies, which were included in each meta-analysis, European experiments needed to be present. We decided on a European focus, as no quality assessment of such studies is present so far and our source of funding, the EJP SOIL, is centered on European research. Out of 386 articles found in literature search, 355 were excluded, as they did not align with our inclusion criteria (see Figure 2 and Table 2).

L 166. "The 17 quality criteria were structured according to three groups"

- How did you choose which criteria among the numerous references used to keep as relevant in your analyses and the ones redundant or not useful?

Due to the year-long expertise of Koricheva and Gurevitch, and the closest connection of ecological meta-analyses to soil and agricultural ones (compared to medicine or social sciences), we based the criteria set on their (2014) "Checklist of quality criteria for meta-analysis for research synthesis, peer reviewers and editors". We removed/adapted criteria so they are more suitable for soil and agricultural meta-analyses, based on findings of related reviews and our own experience when looking at the weaknesses of soil and agricultural meta-analyses published in the recent years. To list some of the changes we made to Koricheva and Gurevitch (2014): e.g., removed criteria 11, included C and T description, recommended log RR as effect size and 1/variance as weighting, added criteria on why parameters for response variable calculation (as SOC) need to be calculated and not

estimates (not present in any criteria set but our own concern), etc. We further added direct quotes and references from other guidelines to describe the criteria in more detail (see Section 2.1. and column “References” in Table 1).

Reference:

- Koricheva, J. and Gurevitch, J.: Uses and misuses of meta-analysis in plant ecology, J. Ecol., 102, 828–844, 708 https://doi.org/10.1111/1365-2745.12224, 2014.

Table 1. Quality criteria 6.

- Could you elaborate on why you consider log ratios to be 'better' effect sizes than hedge's g? Numerous advantages of hedge's g also exist (see for example Borenstein’s book). Numerous meta-analyses in ecology use this specific metric.

We described this later in line 421-427 (here we show an updated version of the lines): “Among the several possible choices in effect size metrics, we recommend using log response ratios when creating soil and agricultural meta-analyses. They are easy to interpret, and effect sizes are not affected by different variances of control and experimental groups. Overall, they are more suitable for meta-analyses studying agricultural management effects on soil parameters as e.g., SOC, than the standardized mean difference (Hedge’s d). When using the standardized mean difference, the results are more difficult to interpret (especially for policy makers or farmers) compared to log response ratios, which can be back-transformed to percent changes from the control”.

In Sect. 3.3 “Results and database presentation”, we mentioned that, in our opinion, raw mean difference (also called

- What do you call “non standard metrics”? Depending on the specific problem analyzed in the meta-analysis, specific type of effect-size could be used (proportion, correlations, ….). These metrics could be the most suitable for some particular meta-analyses, and thus and they should not receive a low-quality rating for this criterion, I think.

By “non-standard metrics”, we talk about studies using other methods of quantitatively synthesizing primary studies than effect sizes as defined by Borenstein et al. (2021). If studies use such methods, they should not name their studies “meta-analyses” or claim that they calculated “effect sizes”, as these terms are specific to the meta-analytical method (Koricheva and Gurevitch, 2014; Borenstein et al., 2021).

Table 1 Criteria 7. Standard deviation extracted >> from each study =2.

- some result papers do not present a dispersion indicator (e.g. SD, CIs,..). Are you suggesting that these papers should be excluded from the database (for the meta-analysis to get 2 points for this criterion)? Moreover, techniques for imputing missing variances are available, and could present an interesting alternative when some papers do not present a dispersion indicator.

Yes, we suggest that only studies which present 1) dispersion indicators as SD, SE or variance or 2) present enough data that allow the calculation of SD should be included in a meta-analysis, as only then weighting by the inverse of variance is possible. We recommend the computation tool by Acutis et al. (2022).

References:

- Acutis, M., Tadiello, T., Perego, A., di Guardo, A., Schillaci, C., and Valkama, E.: EX-TRACT: An excel tool for the estimation of standard deviations from published articles, Environ. Model. Softw., 147, 105236, 628 https://doi.org/10.1016/j.envsoft.2021.105236, 2022.

Table 1 Criteria 8.

- Can you explain how a meta-analysis weights only part of the studies by their variance?
- Furthermore, how do you consider studies that weight the effect sizes by the sample size (number of data)? See for example doi:10.1177/0013164409344534

Some MA did weigh correctly by 1/variance but introduced bias when including studies which did not report SD or SE, as they did not back-transform SD or SE with given data/statistics but only estimated them. E.g., Feng et al. (2020, p. 4): "For studies in which standard deviation or standard error is not available, we assigned a standard deviation equal to one-tenth of the mean, in keeping with the methods of other studies in this domain [47–50]".

We consider MA that weight the effect sizes by the sample size as incorrect, as sample size is not a good indicator for data precision (Hungate et al. 2009). Without including SD from mean when calculating study precision and thereafter the weight of the study (by 1/variance), the meta-analysis was not conducted correctly.

- Feng, An, Chen, Wang, Can deep tillage enhance carbon sequestration in soils? A meta-analysis towards GHG mitigation and sustainable agricultural management, Renewable and Sustainable Energy Reviews, Volume 133, 2020, 110293, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2020.110293.

- Hungate, B. A., van Groenigen, K. J., Six, J., Jastrow, J. D., Luo, Y., de Graaff, M. A., van Kessel, C., and Osenberg, C. W.: Assessing the effect of elevated carbon dioxide on soil carbon: A comparison of four meta-analyses, Glob. Change Biol., 15, 2020–2034, https://doi.org/10.1111/j.1365-2486.2009.01866.x, 2009.

Table 1 Criteria 11.

- Is using Excel for a meta-analysis intrinsically suboptimal? Maybe, it exists some macros or other means to apply all the specifics of meta-analysis models (random forest, proper calculation of tau, …)? In my opinion, this criterion is not precise enough, even confusing. Perhaps you are only focusing on the type of model used in the meta-analyses? (Or maybe analysing if the method is transparent and

reproducible? in which case it is useful to have the type of software used and the codes)

We agree with this passage in Schmid et al. (2013, p. 174): “In the past, spreadsheets have often been used to carry out many of the meta-analyses published in the scientific literature, but we advise that these should no longer be used in research. This is because of the likelihood of programming and transcriptional errors, and because many of the statistical and graphical analyses that have now become standard are not available using spreadsheet analyses, though they are available elsewhere.”

- Schmid, C. H., Stewart, G. B., Rothstein, H. R., Lajeunesse, M. J., and Gurevitch, J.: Software for statistical Meta-analysis, in: Handbook of Meta-analysis in Ecology and Evolution, edited by: Koricheva, J., Gurevitch, J., and Mengersen, K., Princeton University Press, Princeton, 147–194, 2013.

Table 1 Criteria 12.

- How do you consider here the three levels random-effect meta-analytical models that allow for effect-size dependence within a study?

We consider a conventional meta-analysis according to Handbooks of meta-analysis (e.g., Borenstein et al, 2009), but not a multilevel meta-analysis, which is very much advanced method. Moreover, none of 31 meta-analyses, evaluated in our review, applied multilevel meta-analysis, indicating rare use of this complicated method in soil and agricultural research.

Table 1 Criteria 14.

- How do you consider meta-analyses that do test for publication bias and find one: do they get a lower score because the results need to be interpreted carefully, or do they get a higher score because they tested for publication bias?

We assigned a score of 1, when publication bias was tested. The meta-analyses that tested for publication bias mostly did not detect one. E.g.,

1. Bai 2019: “The publication bias analysis suggested that most results in this study are robust (Table S3)”.
2. Feng 2020: “The fail-safe number for publication bias analysis is 45,508, indicating that most of the results considered in this study were robust (Table S2)”
3. García-Palacios 2017: results of bias test not stated
4. Haddaway 2017: “Sensitivity analyses for critical appraisal category and variability type demonstrated no evidence of bias, and there was no evidence of publication bias, with two studies exerting high influence on the model (see Additional file 10).”
5. Mondal 2020: “Publication bias was assessed through histograms (Rosenberg et al., 2000), and in none of the cases, effect sizes showed preference towards positive or negative bias.”
6. Han 2016: “Publication bias was evaluated using the Kendall’s tau rank and Spearman rank48−50 (Table S4). It should be noted that publication bias was detected under

CFM treatment in cool temperate and in the 0–10 yr subgroup. Therefore, such estimates should be used cautiously when applying them to estimate the real soil C sequestration"

7. Li 2020: Supplementary material - not detected
8. Liu 2016: "Both Rosenthal's method and Orwin's method suggested no publication bias across the soil textures on soil CO2 fluxes, SOC, and MBC."
9. McDaniel 2014: results of bias test not stated

L 206. "The results were compared with the meta-analyses identified by Bolinder et al. (2020), who synthesized meta-analyses studying the effects of several management practices on SOC changes in agroecosystems"

- See also maybe :
- Young, M. D., Ros, G. H. & de Vries, W. Impacts of agronomic measures on crop, soil, and environmental indicators: A review and synthesis of meta-analysis. Ecosyst. Environ. 319, 107551 (2021).
- Lessmann, M., Ros, G. H., Young, M. D. & Vries, W. Global variation in soil carbon sequestration potential through improved cropland management. Change Biol. 28, 1162–1177 (2022).

We appreciate the provided literature. Our search was conducted on January 5th, 2021 (no studies published after this point included) and compared with Bolinder et al. (2020). This is why we did not compare our results with reviews published after this date but already existing ones.

L 224. "The aim was to assess how many meta-analyses were conducted on a certain management practice and whether their quality was sufficient to stop the production of new meta-analyses on the respective practice"

- Quality is only one of the criteria for judging the value of continuing the synthesis efforts. A low precision of the effect estimated globally, or new sub-group analyses as a function of co-variate or geographical region, or in combination of other factors, can in my opinion justify the interest of continuing the synthesis efforts in a field (or on a particular practice), even if meta-analyses already exist.

We completely agree with this statement. In our results and discussion section we acknowledged quality and geographical region when stating whether available meta-analyses on a management practice are "sufficient" at the moment.

Results:

L 254. "Scores also experienced a rise (15-year period) and related with the publication year (y= -1889.8980+0.9437*x; R2= 0.39)"

- Is the slope significant? Do you also include "pseudo-meta-analyses" in this analysis?
- It is interesting to note that similar trends have been observed in others papers, see for example El-Rabbany et al 2017, Jamshidi et al 2018, Beillouin, 2019 (Figure S3).

Yes, the regression is significant.
We included all 31 analyzed meta-analyses, which means that also “pseudo-meta-analyses” are part of this regression. Thank you for providing relevant literature.

L 279. “The “Meta-analysis” group consisted of nine quality criteria, “

- the term meta-analysis may not be the most appropriate to deal with a sub-part of the criteria (all criteria are linked to meta-analyses). Replace by statistical analysis? or ..?

As the term refers to the statistical procedure, we decided to use it for the statistical group of criteria.

L 288. “Nevertheless, we urge authors to extract SDs for each study and further weight them by the inverse of variance in order to conduct a high-quality meta-analysis”

- This information is not always presented in experimental studies. (see my comment above)

As we mentioned earlier in the comment responses, we urge to either calculate SD when possible or exclude the study. Otherwise, SD is only estimated when using 1/variance for weighting or no correct weighting is possible, and authors then use incorrect weighing by e.g., number of studies (n). In the Discussion we encourage using an EX-TRACT tool (Acutis et al., 2022) to obtain SDs from available statistics.

L 300. “Only about 25% of meta-analyses had no problems with non-independence of effect size, while the rest extracted several effect sizes per study

This this not always a problem, depending on the statistics made (see for example three levels meta-analyses, meta-analyses with variance-co-variance matrix – Lajeunesse, 2011, …).

Indeed, this method can be used for calculation of aggregate effect size for each study, but, in our opinion, this method has some limitations, as (1) additional knowledge on correlation between multiple outcomes in a study is needed, (2) confounding factors for testing moderators. For example, if in a study there are two treatments - compost and slurry - used with the different N rates and C/N ratio (moderators), after calculating an aggregate effect size, it is not possible to investigate the effect of moderators on outcome, that reduces the quality of agricultural meta-analysis. We changed this sentence as “Only about 25% of meta-analyses accounted for non-independence among effect sizes, while the rest failed to do so”.

Figure 6.

- The "real" meta-analyses (according to your criteria), therefore have lower quality than the average of the others, is that right? Is your criterion relevant, then?

No, this figure actually should show that when considering only meta-analyses that fulfill the cut-off criteria (=true meta-analyses; striped bars), a lower % of studies complies with the

quality criteria 9-17. Simply because the number of MA is lower when excluding “pseudo” MA. The meta-analyses that passed the cut-off criteria and are then assessed for the rest of the criteria are displayed as a percentage of the total number of meta-analyses (n=31). We will change the figures 6 and 7 into stacked bars, which should make them easier to understand.

Discussion:

L 397. “A quality criterium, which is of special significance to the soil and agricultural field, is the inclusion of grey literature”

- Again, I am not totally convinced that this issue is specific to the soil and agricultural field.

Agricultural Research Institutes and Research Stations have local, unpublished reports that can have a value for a research synthesis. For example, several meta-analyses summarized the results of dozens of unpublished reports and hundreds of experiments on nitrogen and phosphorus fertilization.

1. Valkama E, Salo T, Esala M, Turtola E (2013) Nitrogen balances and yields of spring cereals affected by nitrogen fertilization in northern conditions: A meta-analysis. Agriculture, Ecosystems and Environment 164: 1– 13. 10.1016/j.agee.2012.09.010;
2. Valkama E, Uusitalo R, Ylivainio K, Virkajärvi P, Turtola E. (2009) Phosphorus fertilization: a meta-analysis of 80 years research in Finland. Agriculture, Ecosystems and Environment 130: 75–85. https://doi.org/10.1016/j.agee.2008.12.004

L 451. “Therefore, common- and random-effects models are not useable, leading to difficulties in assessing heterogeneity (Gurevitch et al., 2018).”

- I am not sure to understand this sentence? You suggest that when not weighted at all, we can neither use common-effect models nor random-effects models?
- What are the different type of function used to weight studies?

From Gurevitch et al., 2018, p. 179: “Meta-analyses that are not weighted by inverse variances are common and often poorly justified, and present different problems. Unweighted meta-analyses can be unbiased and may provide information on the magnitude of the effects. However, in an unweighted analysis, within- and between-study variation cannot be readily separated, and so common- and random-effects models cannot be used and heterogeneity may be difficult to assess properly. Unweighted meta-analysis also increases the influence of small studies, which have often been found to report larger and more variable effects than those reported for larger studies (as a result of the smaller studies being more likely to suffer from random noise, and possibly publication bias)”.

L 461. “Effect sizes might show a certain amount of variability that cannot be explained by sampling errors alone,”

- Depend on the assumption of the model used. Fixed-effect models consider that the measured effect-size differ to its 'true effect size' (and those of all others studies) only because of sampling errors.

Yes, exactly. This is why we argue not to use the fixed effect model for soil and agricultural meta-analyses. We will adapt the sentence to be more concrete.

"Therefore, we suggest that the topic is well covered for the moment and no further global meta-analysis is needed until there is a substantial number of new results"

- Not agree. Some interesting new meta-analysis results could be produced, e.g. with new method for analyzing the data or new questions. For ex. https://doi.org/10.1038/s41558-021-01075-w

Our statement is concerning only meta-analyses (no other methods, such as the machine-learning algorithm or process-based modeling, which definitely can be useful) on no-till vs. conventional tillage on total SOC only. We agree that when incorporating new methodologies (not meta-analysis) or different research questions (not effects on total SOC but e.g., SOC pools), new synthesis work is of great interest.

L 577. "Quality assessment of meta-analyses, especially in the complex agricultural set up, are highly warranted to harness the power of meta-analyses"

- Does "power" refer to statistical power? Or do you mean potential biased results?

Removing biased results and producing high quality meta-analyses are needed to capture a large set of existing primary data to make sustainable soil management or policy decisions. Removing bias also includes alleviating potential statistical errors and, in future, the reuse data and meta-analysis of large-scale data from multiple independent studies can increase the statistical power to obtain new and robust soil health conservation / soil health policy insights, compared with the analysis of any one study.

L 578. "We demonstrate that meta-analyses in soil and agricultural research encounter specific issues, which differ to other fields like medicine, environment or ecology"

- Not totally convinced by this part.

We changed it to "Experimental studies in soil and agricultural sciences may encounter specific issues, for example, a large variation of pedo-climatic factors, complicated combinations of management practices and the diversity of cultivated crop species cause considerable variation in outcomes". In terms of SOC, the outcomes in the original articles often report stocks for multiple correlated sub-layers that pose problems with non-independence of effect sizes, or with no measurements of bulk density for SOC stock computation, or with missing SDs.

Marco Acutis, 30 Sep 2022

In my opinion the manuscript " Quality Assessment of Meta-Analyses on Soil Organic Carbon" is a high valuable work and can improve the quality of future meta-analytic works. The manuscript clearly highlights what are the main criteria that could be used to obtain a reliable meta-analysis in a group of disciplines (soil and agricultural sciences. So, the idea to create a framework of criteria to be met to obtain a substantially correct meta-analysis, also in consideration that in soil and agricultural science (as reported by the Authors of this manuscript), the introduction of the meta-analysis is relatively new. In consequence, frequently the principles and the methodology to apply this kind of analysis there are not well known, and this work could be very useful not only for the scientific community.

Moreover, the manuscript gives a clear definition of meta-analysis from operational point of view, starting from search of primary studies up to the definition of correct statistical method and what - really important- need to be made available to the scientific community

So, I hope that this work can contribute to produce more reliable meta-analysis in soil science and agricultural field, more and more interesting and useful for policy makers and also for public opinion.

May be interesting to introduce some short consideration about the minimum number of primary studies needed for a meta-analysis and an additional remark about the fact that frequently meta-analysis with a large number of "points" (I use the term "point" because in this case they are not independent studies) often don't meet the independence requirement.

Also a consideration about the use of sample size as weight could be interesting, even if I agree with the authors that it is not a good choice (and it is acceptable only for fixed model)

In fig. 3B seem that a black dot is missing for the year 2014, there is only the number "19" and not the related dot.

As a final evaluation for sure I recommend the publication of this manuscript, with minor revision.

The authors are thankful for this very positive feedback and the presented ideas on how to improve the manuscript. We will include the recommendation for a minimum number of studies to be incorporated and emphasize the issue of non-independence when several observations are extracted from one study. We strongly believe that weighting by sample size is incorrect, as it does not acknowledge the precision of the study outcome (SD). Figure 3B will be adapted according to your observation.