Dear editor,

We very much appreciated your encouraging and insightful comments and those of the referees. We have endeavoured to respond to all suggestions and comments, which further improved the understanding and potential impact of our manuscript. Detailed responses are given below (in blue). In case of further queries, we are happy to clarify any further details and look forward to your reply.

Sincerely,
Sara Vicca, on behalf of all co-authors

Responses to the comments:

EDITOR

The reviewers have submitted their reports and all three acknowledge the relevance of this paper and its suitability for the special number of BG. The three were positive about publishing the paper, after providing clarifications on a number of issues. Please, consider their reports and answer their comments, indicating clearly whether you accept them or not, and the reasons for doing so. Based on the reviewers reports and on my own reading of the paper, following are some of the main points that need to be addressed in the next round:

1. Strength of the approach
The reviewers found as an important asset of the paper the procedure used to assess whether past experiments modifying rainfall can help us understand future climate change impacts on soil CO2 efflux (SCE). This is a strong point that merits being highlighted. Elaborating in the discussion on the pros and cons of this approach, and comparing with similar attempts on other processes of relevance aiming at extrapolating current experiments into a future climate is worth highlighting in the discussion.

We agree with this suggestion and added a new section to the discussion (Section 4.4 in the revised manuscript) in which the new approach is discussed. This Section reads as follows:

4.4 A novel approach revealing limitations of current experiments and recommendations for future experiments

At present, inter-site comparison of effects of altered precipitation is seriously hampered by the lack of data necessary to quantify the treatment as experienced by the biota (i.e., the actual treatment; Vicca et al., 2012). Without such data, conventional meta-analytical analysis of cross-experiment variation in ecosystem responses to precipitation manipulation is prone to artefacts related to the enormous variation in the actual treatment; magnitude, timing and duration of drought and rain events vary substantially among experiments, and soil type and rooting depth considerably influence the way plants and microbes experience a treatment (Vicca et al., 2012). The novel approach presented here was developed specifically to avoid these problems. This is accomplished by analysing within-experiment responses (through calculation of a predictability index) prior to across-experiment comparison (via CART analysis). Although also with this method, treatments remain largely incomparable (hence, if cross-experiment differences would occur, these could be due either to variation in the actual treatment or to differences in ecosystem response, or a combination), our method does provide mechanistic insight into the responses to altered precipitation. Importantly, the results are less prone to the large variation in the actual treatment. It would be particularly interesting to combine this approach with a quantification of the actual treatment such that moisture responses of SCE in various ecosystems can be elucidated.

The approach used in this study fully exploits the potential of the available datasets by taking advantage of the multiple measurements of SCE made in each experiment. However, this method is applicable only when sufficient data are available (we discarded six experiments with ≤ 10 data points), and when a reliable model can be fitted to the control data (in our study, seven experiments were discarded because of the poor quality of the model fit through the control data; Fig. 2). For this reason, and because the CART analysis suggests that frequent SCE measurements are essential to detect deviations of the moisture response of SCE in the treatment as compared to the control, we recommend that future experiments that aim to test the response of SCE to altered precipitation seek to obtain high-frequency SCE measurements. This recommendation also applies to variables such as photosynthesis and ecosystem respiration that can be measured at high frequency with automated cuvettes and are therefore suitable for testing as in this study.

2. Highlighting the limitations of our current experiments

An issue that needs to be given further consideration is the limitations that our current experiments have for realistically mimicking future conditions, and how much our experiments thus limit further extrapolations. Water manipulations are carried out in a context of current non-rainfall climate (i.e., drought treatments are implemented in temperature conditions that may not coincide if a true drought had occurred; probably, something similar could be said about irrigation experiments). This may or may not affect the results of the experiments used in this paper. Whether this is relevant or not for the process under consideration is something that authors should further mention and discuss.

We agree that under climate change, changes in precipitation are likely associated with temperature increases, and interactive effects can further complicate the ultimate response of SCE and of the entire ecosystem. Testing multifactorial effects is thus crucial for improving our understanding and for reliably predicting responses when climate changes. Nonetheless, process-understanding begins with elucidating the single factor responses. In our study, we point out that this is already problematic for precipitation manipulation experiments. In the revised manuscript, we dealt with this issue in the concluding remarks, where the following paragraph was added:

Using single-factor experiments, this study demonstrated that current relationships between SCE and soil moisture should not be extrapolated to predict SCE when precipitation patterns change. However, climate change not only involves changes in precipitation regimes, but also other global change factors. Droughts are often associated with warm periods or heatwaves, and in combination with a heatwave, drought effects are typically exacerbated (Reichstein et al., 2013). This implies that thresholds for structural changes in the ecosystem may be passed earlier, which most likely makes predictions based on current-climate observations even less reliable than our analysis may suggest.

3. On the representativeness of current experiments

The reality of extant experiments is probably far from being optimal for the question of concern. There are biases in the selection of biomes, of sites within biomes, and of vegetation-types within sites (e.g., grasslands are over represented with respect to the rest of vegetation types), among other. In section 4.1., when initially discussing the results of the paper, it is important to indicate how biased current experiments are in terms of biomes and vegetation types within biomes. Really and truly, the available dataset is not representative of extant vegetation in the world.

In response to this comment, we indicated that our dataset is restricted mostly to temperate- and subtropical regions. We do not fully agree on the bias towards grasslands though. This apparent bias disappears when considering site level instead of experimental level. Our dataset includes 11 forest sites (for a total of 13 forest experiments), 9 shrubland sites (11 experiments), and 7 grassland sites. The latter are, however, represented by 31 experiments. Because the CART analysis weighs the multiple experiments belonging to a single site, our statistical approach further reduces any site bias. We did therefore not go into detail about the representation of the different vegetation types in our dataset, but we did indicate that we found only one experiment in an agricultural field and have only one hydric site (see Sections 3.1 and 4.1 of the revised manuscript).

4. Treating experiments at the same site as independent data points
The approach followed uses all available experiments as if they were independent from one another, even if several sites had 2 or up to 19 experiments. The assumption of independence of these experiments, and its consequences for the approach used, needs to be considered. For instance, even if no statistical testing was done, in 31 experiments out of the 38 considered, H1 was not rejected. This is a strong result that is interpreted without needing any additional testing. How would these figures change if, instead of experiments, sites (with repeated measures within sites for experiments and vegetation types) would have been considered? Since this paper has a strong methodological component, this is something that merits attention.

In an analysis like this, it is indeed important to take into account the fact that several sites are represented by more than one experiment. For that reason, the CART analysis included as a weight factor the reciprocal of the number of experiments for each site. Hence, the CART analysis in the previous version already showed the requested results, but admittedly, this was not sufficiently emphasized in the results and discussion. In the revised manuscript, we explicitly mention in the results section that the CART analysis takes into account the dependency of the results from different experiments within one site and also indicate the number of independent sites for which H1 was rejected and for which not (Section 3.2, last paragraph). In the revised manuscript, we also repeat this in the legend of Fig. 4, which shows the results of the CART analyses.

5. On the limited discussion of your results
SCE is an important process, and answering the question you posed is utmost relevant. The reviewers felt, and I agree with them, that after doing your analysis you did not rightly discussed your results, as one would expect from this exercise. It is very important to further delve into this point in the discussion, and answer your question, and the limitations of your answer in a much clearer way. Additionally, much of the discussion went to the details of some experiments. The important point in those cases is finding the generalities or the particular processes that may contribute to support or not support your conclusion, or to help guiding future research. This is not fully captured in your discussion and it needs further focus.

We agree with the editor and reviewers. In response to this comment, we added focus to section 4.3, where the individual responses of the sites with daily data are discussed. We added in the first paragraph of this section a sentence to indicate the purpose of briefly discussing the observations at individual sites. We further reduced the explanation of the Birch effect to improve the focus of the results from Solling in view of our study. The paragraph of each of the discussed experiments now ends with a sentence indicating the need for data that was indicated by its results.
In the revised manuscript, we added Section 4.4 where the pros and cons of the approach and its overall outcome are discussed. Also the conclusion section is completely rewritten to clearly answer the question posed in the title.

6. Many experiments did not even pass an initial scrutiny
In the discussion, before going to the CART, something needs to be said about the fact that, before an analysis could be run, out of 58 cases only 38 could be used, that is only 65% of experiments were proper to further analysis. This is something that needs to be discussed as well. Discarding 35% of the studies after your selection of the ones you thought were appropriate is not a minor issue.

We agree with the editor and emphasized this point in Section 4.4 of the revised manuscript, where we discuss the pros and cons of our approach.

7. Defend your hypothesis
You need to believe your hypothesis and defend it, whether it is verified or not. If not, withdraw the hypothesis. You ignore the fact that, out of 38 experiments, the hypothesis could not be rejected in 31, and only in 7 was rejected. It is difficult to accept that you find that 37 out of 38 experiments go along with your reasoning, and later you basically not discuss this because you found that measurement

frequency is the relevant variable. The apparent contradiction that the hypothesis is accepted when the frequency of available data is low and that it is rejected when data frequency is high needs further elaboration. If that is the case, you need to enter into the details of why this is possible. How is it possible that such a relevant, but that one could consider a priori trivial, finding went unattended? You did not discuss at all this point. This is important in a perspective of future experimentation in order to not continue doing the wrong things, should this be the case.

In the philosophy that a hypothesis should be falsifiable, we opted to test the hypothesis that the relationship between SCE and ST and SWC can be extrapolated. Based on the principles explained in the introduction, we did expect falsifications to occur. It is also very important to realize that not rejecting the hypothesis does not necessarily imply that the relationships between SCE and ST and SWC predict well SCE under altered precipitation regimes. This is now better clarified in the last paragraph of the introduction, where the following sentence was included:

Based on the above-mentioned mechanisms, we expect H1 to be rejected, not only when SWC in the treatment exceeds SWC in the control, but also after SWC recovered but other ecosystem properties did not.

8. Modeling soil moisture content vs. measuring it
This was a well thought approach. However, you virtually ignored it in the results, discussions and conclusions. Actually, the presentation of the results is misleading. There is potentially a lot to learn from this exercise, positive or negative. Just going to the negative, if we cannot use this type of modelling, you need to put your findings in context, because this is a relevant result. This part deserves the consideration that you gave it when planning the experiment. So far, is missing, and it should not.

We agree with the editor and have rewritten Section 3.2, where we now pay attention to the results of the bucket model approach:

In order to test whether artefacts related to SWC measurements were responsible for rejecting H1, we replaced SWC in model 4 by the bucket model results. This exercise provided results of acceptable quality (i.e. normal distribution of the residuals and an $R^2 \geq 0.30$; see SI) only for 16 of the 45 experiments, indicating the limitations of this approach. Nonetheless, for 14 of these 16 experiments, the outcome of the bucket model approach was the same as the outcome of the SWC approach. Importantly, rejection of H1 was confirmed for three of the seven experiments (i.e., for Solling, Stubai and TurkeyPoint; Table S5). For the other four experiments where H1 was rejected using the SWC approach, the low quality of the fits based on the bucket model approach did not allow this test. In any case, the bucket model approach indicates that artefacts related to SWC measurements are unlikely responsible for rejecting H1.

9. The reviewers indicated other minor points, such as better reconciliation of the abstract with the results and conclusions, clarification on the conclusions, and additional details on the sites used for each purpose. The data are there, but being a dense paper, any effort to improve the readability and help the reader understand what you did will be greatly appreciated.

The conclusions were completely rewritten. The abstract and conclusion section now better match and provide a clear answer to the question posed in the title.

Other minor comments include:
10. Page 7, lines 11-14 (Extremes ... 2012): This is a rather limited presentation of changes rainfall-related extreme events in the world. Please, modify it to be more comprehensive, even if this not the focus of your paper.

We agree with the editor that it would be interesting to provide more details about projected changes in precipitation patterns. This is difficult, however, because these model projections are still very uncertain. In the revised manuscript, we indicated this uncertainty as well as the projections for which models are quite consistent (Section 1.2 in the revised manuscript):

Extreme events such as severe heatwaves and droughts are expected to increase in intensity and periodicity. Although current model projections of climate extremes remain uncertain (with contradicting results from different models), consensus is growing that for example in the drier temperate regions the number of consecutive dry days will increase (Orlowsky and Seneviratne, 2012; Seneviratne et al., 2012). In the Mediterranean region longer dry spells and more intense precipitation events are very likely (Seneviratne et al., 2012).

11. Page 8, lines 13- 26. (We... factor): The figures and tables are presented in a rather haphazardly way. Please, organize your text so that the figures and tables in both the main text and appendices as well as in the supplementary material are presented correlatively and in order. Following is their listing as they were called in: Figures: Fig S2, Fig S3, Fig. 1; Fig. 2, Fig. S1; Fig. S2; Fig 1; Fig. 1; Fig. 3a; Fig. 3c; Fig. 3d; Fig. 4; Fig. S3; Fig. S3; Fig. 5; Fig. S3; Fig. 5; Fig. 5; Fig. S2; Fig S2. Note that Fig. 3b has not been called. Tables: Table C1; Table A1; Table S4; Table C1; Table C1; Table A1; Table B1; Table C1; Table B1; Table B1; Table C1; Table C1; Table C1, Table C1; Table B1; Table B1, Table C1; Table B1; Table C1; Table C1; Table C1; Table B1; Table C1. Note that Tables S1 to S3 have not been called in, except generally as SI.

Thanks for bringing this to our attention. In the revised manuscript, the figures and tables are listed in the order of appearance in the manuscript. This required changing the order of appearance of Tables and Figures in Supplementary Information.

12. Page 9, line 3-6: SR is not defined. Please, use consistently SCE, here and throughout the text and appendices and supplementary material.

Thanks for pointing out this inconsistency. This is corrected in the revised version of the manuscript.

13. Page 9, line 12: call out Table S2 and Table S3.

Done. In the revised manuscript, the previous Table S2 and S3 are Table S3 and S4, respectively. On p. 9, they are referred to.

14. Page 9, line 15: on the use of model 4. While this is reasonable, it is unclear to what extent model selection (one for all vs. the best for each) affected your results. This is something that needs to be considered.

In the last paragraph of p. 9 of the revised manuscript we now mention that the results were generally very similar when using one of the other models (although fewer experiments could be used because less sites passed the criteria of normal distribution).

15. Page 10, line 4: Edit (experiments at Solling and...

Done.

16. Page 11, line 8-9: SR is not defined. Use SCE consistently, here and elsewhere.

Done.

17. Page 12, last line: a minor comment declaring the bias in the type of vegetation within biome is also appropriate.

Done.

18. Page 13, line 10: Edit (Fig. 13b to 13d) to call all panels in Fig. 13.

I suppose this refers to Fig. 5. I edited the text as suggested, including the panel number when referring to Fig. 5.

19. Page 13, line 14: Edit (4 instead of four), for consistency.

It is unclear what the editor refers to. There was no 'four' in the indicated line. We followed the standard instructions to write in full all numbers < 10.

20. Page 13, line 15: Edit (indicate the column name [Robust?] to help the reader).

Done.

21. Page 13, line 16: you indicate "results for individual experiments were confirmed when SWC in model 4 was replaced by the bucket model results...". However, the results [Robust?] were coincident in 7 out of 37! I am not sure that this sentence captures your results. This needs further clarification. Please, call out the specific table you are referring to in the SI (Table S5) and add a column in this table to signify whether the results from this exercise were coincident or not with the results based on the actual measurements in the field.

We rephrased this sentence in the results section, indicating also the limitations of the bucket model approach (as requested in an earlier comment – see point 8 above). Specifically, the following paragraph was added:

In order to test whether artefacts related to SWC measurements were responsible for rejecting H1, we replaced SWC in model 4 by the bucket model results. This exercise provided results of acceptable quality (i.e. normal distribution of the residuals and an $R^2 \geq 0.30$; see SI) only for 16 of the 45 experiments, indicating the limitations of this approach. Nonetheless, for 14 of these 16 experiments, the outcome of the bucket model approach was the same as the outcome of the SWC approach. Importantly, rejection of H1 was confirmed for three of the seven experiments (i.e., for Solling, Stubai and TurkeyPoint; Table S5). For the other four experiments where H1 was rejected using the SWC approach, the low quality of the fits based on the bucket model approach did not allow this test. In any case, the bucket model approach indicates that artefacts related to SWC measurements are unlikely responsible for rejecting H1.

The match between the SWC approach and the bucket model approach is better clarified also in the legend of Table S5 and italic fonts indicate the experiments where the results did not match.
The legend of Table S5 includes this clarification: *For comparison, 'Robust H SWC' indicates whether results based on SWC measurements were robust or not (as given in Table C1). Only for 16 experiments was this comparison possible. For two of these the bucket model approach did not correspond to the approach using SWC measurement (indicated in italics).*

22. Page 13, line 19: Edit (7 instead of seven), for consistency.

It is not clear what the editor refers to. We followed the standard instructions to write in full all numbers < 10.

23. Page 13, line 4 (starting from the bottom) to page 14, line 3: When Pi ... in SCEtreatment.":
This is a clarification that belongs to the methods section.

This explanation is also given in the methods section (below equation 5). In order to facilitate the understanding of our approach and the comprehension of our study, this clarification was repeated in the results section. Because we think this is helpful for the reader, we kept it there.

24. Page 14, line 7: Experiments with daily measurements showed a significant trend. Please, bring to the discussion and, eventually, analyze whether the number of measurements could have influenced your results.

The number of measurements was included as a predictor variable in the CART analysis. This is indicated in the methods section (p. 11 in the revised manuscript). CART analysis did not indicate the number of measurements as an important predictor variable.

25. Page 14, 2nd paragraph, and until the end of the results section: This text is more relevant in the discussion (Section 4.3): Why you did or did not find a trend, and the generalities that emerged from this, etc.

This comment refers to where we briefly go into the results of specific experiments (presented in Fig. 5). Here we disagree with the editor and think that the information in Section 3.3 is particularly useful for guiding the reader through the patterns observed for the five experiments with high-frequency data. To make clear the reason why we elaborate about these patterns, we added the following sentence to the paragraph just before the dissection of the individual sites: *These patterns can expose the underlying reasons for rejecting H1*.
We hope the editor can agree on keeping this text in the results section.

26. In the supplementary material, you are calling for Table 2 on four occasions, but there is no longer a Table 2. Please, check the numbering of figures and tables in the main text and supplementary material for correctness.

Thanks for noticing. This is corrected now (referring to Table C1).

27. Table S1. It is unclear what is meant by "(dominant)". Please clarify the text to make unambiguous. Additionally, BG does not have instructions as to how list species and whether full listing, following the nomenclature code, should be followed. Personally, I would suggest that listings are complete or at least a reference to which floras or relevant documents are used. In some entrances (Tolfa sites), the second name of the binomen of the species is given in capital letters (Arborea), but should be lowcase (arborea). Names are recommended to be in italic. Additionally, in some sites (Walker- Branch), only the genus is given (Quercus spp., Acer spp.). In the original publication the dominant species are given. Please refer to it and provide at least the most relevant species.

Thanks for pointing out this error. We made corrections as suggested and changed 'dominant' into 'the most abundant' because some ecosystems are very species-rich and it is not possible to list all species (e.g. the tropical forests).

28. Fig S1: Note that in both X and Y axis SR instead of SCE (units) is given. Please, modify both legends in all graphs.

Thanks for bringing this inconsistency to our attention. This is corrected in the revised version, where the previous Fig. S1 became Fig. S3 and indicates SCE (µmol m^{-2} $s^{-}1$) in the axis titles.

29. Fig 5 and S2 and S3: The representation of time in the X axis makes very difficult to follow the course of events. A continuous measure (DOY) would have facilitated reading, plus seasons where these a relevant. Please, consider this in your final submission.

At first sight, we agreed with this suggestion. However, in our case, providing DOY is less convenient for the reader because the datasets for the different sites span very different periods (ranging from <1

year to 11 years). For this reason, we opted for a fixed number of ticks in the X-axis. In order to improve the readability, we altered the representation of the date in the revised version, which now shows mmm/yy of dd/mm/yy, and always begins on 1 January and ends on 31 December.

30. Fig. S2 and S3: small black circles are mentioned, but there are not such circles. Black diamonds are used instead. Please, correct the captions.

In response to this comment, we changed the symbols in the figures, which now truly are small black circles.

31. Caption Table. S2 and S3: "For each experiment (add: with more than 10 data points), the : : :"

Done (Tables S3 and S4 in the revised version).

32. Test for artifacts, line 8: "lilliefors" should be Lilliefors (capital L).

Done.

33. Supplementary information content: Please, number the pages.

Done.

I will be happy to consider the acceptance of your paper after these issues have been satisfactorily addressed, and I look forward to your response.

We once again want to thank the editor for his insightful comments and suggestions that substantially improved our paper.

REFEREE 1

34. Soil CO2 efflux is the second largest CO2 flux and slight changes in the factors influencing soil CO2 efflux are likely to have strong impacts on the global carbon budget. The authors were using the results from rainfall manipulation experiments to test if current responses to soil temperature and soil moisture can be used to predict CO2 efflux under modified precipitations regimes. The analyses and conclusions are based on 31 experiments (with the majority of experiments conducted in the temperate zone). The used approach provides new insights and fits within the scope of the special issue. The paper is well done on the whole.

We thank the referee for this positive assessment.

35. Yet, I would like to ask the authors to address the following issues. Please state more clearly how many experiments were used to test the hypothesis (24 or 31, see Figure 4). Overall it is difficult to follow haw many experiments were excluded for which reason.

We agree with this concern; in the revised version, we have better clarified the number of experiments used at each step. In summary, 45 experiments were used to test the hypothesis whether or not observations from the control plots can be used to predict SCE in the treatment. This test consists of two criteria-tests: (1) normal distribution of residuals and (2) RMSE should not be larger than double RMSE in the control. Only if both criteria give the same result, the outcome is considered robust and can be used for further analyses. We therefore excluded 7 experiments, thus keeping 38 experiments for the next step in the analysis: trend analysis and CART analysis. The CART analysis thus included 38 experiments (14+24 in Fig. 4). This protocol is clarified in Fig. 2 of the manuscript. Of the 38

experiments, H1 was rejected in 7. For the remaining 31 experiments, H1 could not be rejected (24+7 in Fig. 4).
In Section 3.2 of the revised manuscript, the experiments used at each step of the process is better clarified. Note also that Figure 2 summarizes the protocol and includes the number of experiments used at each step (this figure has been further improved in the revised version of the manuscript). Table A1 provides detailed information about the experiments that were included at each step. Because the confusion arose from Figure 4, the legend of this Figure now includes an explicit statement of the number of experiments used.

36. Out of these 31? experiments, 14 are conducted in close vicinity (ThuringerSchiefer). Are the findings influenced by such a strong bias towards one area?

This is a good remark. The CART analysis takes into account the interdependency of results from multiple experiments belonging to the same site. Our findings are thus not influenced by this site being represented by multiple experiments. Admittedly, this was not sufficiently emphasized in the previous version of the manuscript. In the revised manuscript, we explicitly mention in the results section that the CART analysis takes into account the dependency of the results from different experiments within one site and also indicate the number of independent sites for which H1 was rejected and for which not (see Section 3.2 of the revised manuscript).

37. The hypothesis was rejected for the seven experiments which had the most reliable data set (i.e. those providing high-frequency measurements of SCE). In contrast, the authors highlight the importance of high-frequency measurements. Please clarify.

Only with high-frequency measurements, are we able to robustly test whether or not SCE in the treatment can be predicted from the observations in the control. Experiments with a low measurement frequency do not provide the necessary data to robustly test our hypothesis. In response to the various related comments of referees and of the editor, we emphasized this issue in the revised version of the manuscript and especially in Sections 4.4 and 5. Section 4.4 is provided also in the answer to comment 1. Section 5 reads as follows:

Is it possible to extrapolate the relationships between SCE and its abiotic drivers – soil temperature and soil moisture – to predict SCE responses to changes in precipitation patterns? According to our results, the most justified answer to this question is 'no'; although for the majority of the experiments we could not falsify the hypothesis that we can predict SCE under altered precipitation regimes from current-climate observations. As discussed, all experiments with daily SCE measurements (i.e., the experiments with the datasets most reliable for this exercise) revealed that SCE in the altered precipitation treatment could not be predicted from the control observations. We postulate that at least part of the experiments with infrequent measurement schemes provided insufficient capacity to detect shifts in the climate-dependencies of SCE. In other words, crucial patterns in SCE likely went undetected for these experiments. Importantly, the erroneous predictions in the experiments with daily SCE measurements were not related to extrapolation beyond the range for which the model was parameterized. Instead, these experiments provide insights of likely mechanisms (e.g. the Birch effect) that cause SCE in the treatment to deviate from what would be expected from the control observations.

Using single-factor experiments, our study demonstrated that current relationships between SCE and soil moisture should not be extrapolated to predict SCE when precipitation patterns change. However, climate change not only involves changes in precipitation regimes, but also other environmental forcing factors. Droughts are often associated with warm periods or heatwaves, and in combination with a heatwave, drought effects are typically exacerbated (Reichstein et al., 2013). This implies that thresholds for structural changes in the ecosystem may be passed earlier, which most likely makes predictions based on current-climate observations even less reliable than our analysis may suggest.

At present, the available data do not enable full elucidation of the mechanisms that complicate extrapolation of current-climate observations of the moisture response of SCE to predict SCE when rainfall patterns alter, and this likely applies also to other ecosystem and carbon cycle processes. If we are to fully understand ecosystem responses to altered precipitation, we need more experiments establishing response functions across a broader range of precipitation regimes, annual temperatures, soil moisture conditions, and vegetation types (especially in boreal and tropical regions). Such experiments should make accurate measurements of water availability, they should consider both instantaneous responses and the potential legacy effects of climate extremes, and would benefit from a holistic approach that allows elucidation of underlying mechanisms. Future studies should make particular effort to obtain high-frequency measurements, which - as we demonstrated - are essential for capturing dynamic responses during drying and after rewetting, and for quantifying their implications for the carbon cycle in a more extreme climate.

38. The authors tested the "hypothesis (H1) that the relationship between SCE and temperature and volumetric soil water content (SWC) observed over time in the control plots can be predicted to ...". However, temperature was only taken into consideration while testing model 1-4.

Thanks to this comment, we now see the confusion caused by including 'ST' in the hypothesis. To avoid this confusion, the hypothesis was rephrased and now reads as follows:
'H1: hypothesis that the soil moisture response of SCE as observed from fluctuations over time in the control plots can be extrapolated to predict SCE in plots exposed to a different precipitation regime'.

In the methods, we explain that ST hardly differed between control and treatment plots, but needs to be taken into account because variation of SCE over time is strongly related to variation in ST. (first paragraph of p. 9).

39. The authors tested for artefacts related to SWC measurements. What about the effect of the set-up, scale and duration of the different precipitation manipulations experiments on SWC and soil CO2 efflux. How much of the variability is "experiment" dependent? Please elaborate on the statement made in line 2-5 on page 867.

In order to answer this question, the CART-analysis was redone including also the experimental duration and the manipulation technique as predictor variables. This did not change the outcome of the analysis (i.e., neither duration, nor technique was selected as a predictor of whether or not H1 could be rejected). This is clarified in the methods section (first paragraph of p.11, where the predictor variables are indicated).

40. All experiments (except one) are located in the temperate zone. Please re-phrase your statement given in line 25-26 page 866.

Our dataset includes three tropical forests (Caxiuana, SulawesiForest and SulawesiCacao; indicated by the first letter 'A' in Köppen classification), as well as different subtropical and Meditteranean sites (a.o. Aranjuez, Garraf, Prades, Tolfa) that are indicated with Csa, Csb, Cfa and Cwa in the Köppen classification (Table A1; Köppen codes are clarified in Appendix A). Hence, we think that the statement that our dataset covers different climate regions, but is dominated by the temperate zone is justified.

41. The discussion section is mainly addressing site-specific aspects but an answer to the question "can current moisture responses predict soil CO2 efflux under altered precipitation regimes" is largely missing.

We agree with the reviewer and have rewritten the discussion and the concluding remarks to (among others) better answer to the question posed in the title.

REFEREE 2

42. The manuscript addresses a very interesting question, using a very robust and thoughtful methodological approach. Vicca et al. aim to evaluate to what extent the actual relationship between soil moisture and soil CO2 efflux (SCE) can be use to predict the response of SCE to altered rainfall patterns. The paper is topically appropriate for this special issue, and it might make a suitable methodological contribution to be taken into account for future experimental and modelling works in relation to this topic. However, I miss a more direct answer and discussion to the question presented in the title. From the initial 58 experiments considered initially in this work, for one season or another, only data (temperature, moisture and SCE) for 31 of them could be extrapolated to predict SCE in plots exposed to a different precipitation regime. According to these results the answer to the question could be "NO" (at least with data provided by the experiment included in this work). Thus, I would recommend clarify this fact in the abstract and in the discussion section.

We thank the reviewer for this positive assessment and agree with the comment raised. In response to this suggestion, we rewrote the conclusion section to clearly answer the question posed in the title. The new conclusion is provided also in the answer to comment 37.

43. Even though, fortunately this manuscript goes further and also explore the reasons why this extrapolation sometimes cannot be possible or does not work (<10 data points, residuals not normally distributed, not robust results,...). It is interesting the description carried out in this work about how should be the experiments and results obtained with them in order to be able to use the approach presented here and to implement it for further model projections of future climate. Thus, it is very important to mention in this manuscript that not all data of soil CO2 efflux from datasets could be suitable for doing future projections, which it is of critical importance for modelers and the interpretation of their results.

We thank the reviewer for bringing this to our attention. This comment is addressed in Section 4.4 of the revised manuscript, which discusses the approach and the data required for it (see also the answer to comment 1).

44. In addition, I have some additional concerns and I think that some clarifications are still needed:
- Fig 1 includes all the 58 experiment, however from all of them only 31 give robust result in the extrapolation to predict SCE under different precipitation regimes. It would be interesting to discuss a bit about the general response of SCE to precipitation manipulation for these 31 experiments, and the experiments where extrapolation was not successful. Does decreased precipitation typically reduce SCE whereas enhanced precipitation increased SCE? Which climate regions do they represent? In addition, it would be desirable to describe in more detail the type of vegetation that is represented in these experiments. As the main experimental effort has been carried out in grasslands, further experiments need to be done in other natural ecosystems (forest, shrubland, agriculture...)

Here the reviewer seems to suggest that we perform a meta-analysis to test for effects of different climates and vegetation types. However, this is exactly what we do not want to do because we believe that results of a meta-analytical approach are not trustworthy. Because the experiments differ substantially in the applied manipulation, and the actual treatment as experienced by the biota cannot be quantified, a difference between e.g. grasslands and forests as detected via a meta-analysis could be due to either differences in the actual treatment, to differences in the ecosystem response or a combination of both. In response to the reviewer's suggestion, we discussed this issue in the new Section 4.4 of the revised manuscript (see also answer to comment 1).

The reviewer further comments on the bias towards grasslands in our dataset. However, this bias is less obvious when looking at site level instead of experimental level. Our dataset includes 11 different forest sites (for a total of 13 forest experiments), and 9 different shrubland sites (11 experiments), but only seven different grassland sites. The latter are, however, represented by 31 experiments (see also answer to comment 3).

45. According to the results, for experiments with median measurement intervals of SCE larger than 11 days, H1 was never rejected, whereas H1 was rejected for seven of the 14 experiments with intervals _ 11 days, which included all five experiments with daily measurements. In relation to this you also point out that "we have missed important SCE treatment responses in experiments with larger measurement intervals". Thus, we could consider that although in some cases current moisture responses could be used to predict SCE under altered precipitation, this prediction could not be accurate because the infrequent measurement schemes have insufficient capacity to detect shifts
in the climate-dependence of SCE. Taking into account all mention above, these results emphasize the unsuitability of predicting SCE under altered precipitation using current moisture responses; instead of "...the need for high frequency SCE measurements to fully capture the response of SCE to changes in precipitation and other climatic variables such as temperature" (that although it could be true, it is not emphasized for the results obtained in this work).
In addition, it would be desirable to clarify in the discussion the possible reasons why H1 was rejected for 7 experiments (explaining it in general terms instead of going case by case)

In response to this and similar comments, we have improved the discussion and have completely rewritten the conclusion section. This now clearly answers the question posed in the title and generalizes the rejection of H1 for the 7 experiments. The conclusion section is provided in the answer to comment 37.

46. In the discussion section the results from particular cases are commented in a quite deep detail, although there are not representative of any clear general tendency. This fact makes the discussion section hard to follow, and the reader can get lost in the details, forgetting about the main points. Thus, I would suggest to the author to avoid the discussion of particular cases when it is not necessary. Maybe a shorted discussion section but more focus in the main points that need to be mentioned could be much more interesting for the readers.

In response to this comment, we slightly shortened the discussion of individual experiments and ended the paragraph for each experimental site with a sentence that summarizes the need for data revealed by its results (see Section 4.3 of the revised manuscript). The results from these experiments are important because they indicate measurements that are needed to better understand the response of SCE to changes in precipitation.

47. Some other important points that should be included in the discussion or mentioned in more detail could be: a) the effect of rainfall amount as well as timing, frequency, intensity. In relation to it, more experiments manipulating precipitation should be conducted to elucidate the impact of a wider range of possible scenarios; b) the fact that precipitation and temperature effects can have complex interactions in the ecosystem could come to unsuccessful predictions if only single factor experiments are considered. Under a climate change scenario, changes in precipitation regime may have associated changes in temperature. In relation to it, in which extent the experiments considered in this work are realistic? Did all the precipitation manipulation experiment present a range of temperatures close to the climate change projections for these rainfall scenarios? This is also an important fact that needs to be taken into account in order to discuss the results obtained in this manuscript.

a) This comment is addressed in the revised conclusion section, and more specifically in the following sentence:
If we are to fully understand ecosystem responses to altered precipitation, we need more experiments establishing response functions across a broader range of precipitation regimes, annual temperatures, soil moisture conditions, and vegetation types (especially in boreal and tropical regions).
b) We agree with the reviewer that under climate change, changes in precipitation are likely associated with temperature increases, and interactive effects can further complicate the ultimate response of SCE and of the entire ecosystem. Testing multifactorial effects is thus crucial for improving our understanding and for reliably predicting ecosystem responses to

climate change. Nonetheless, process-understanding begins with elucidating the single factor responses. In our study, we point out that this is already problematic for precipitation manipulation experiments. In the revised manuscript, we dealt with this issue in the conclusion section, to which the following paragraph was added:
Using single-factor experiments, our study demonstrated that current relationships between SCE and soil moisture should not be extrapolated to predict SCE when precipitation patterns change. However, climate change not only involves changes in precipitation regimes, but also other environmental forcing factors. Droughts are often associated with warm periods or heatwaves, and in combination with a heatwave, drought effects are typically exacerbated (Reichstein et al., 2013). This implies that thresholds for structural changes in the ecosystem may be passed earlier, which most likely makes predictions based on current-climate observations even less reliable than our analysis may suggest.
.

48. In the conclusion section I would except: a) a more complete list of reasons for which moisture responses from some current experiments cannot be used to predict soil CO2 efflux under altered precipitation regimes and, b) a more complete description of how should be these manipulation experiments in a future if we want to be able to compare their results and use them for future predictions carrying out approaches as the one presented in this manuscript.

a) A complete list of reasons for which moisture responses from current-climate observations cannot be used to predict SCE under altered rainfall patterns would take at least two extra pages and is quite speculative. Because we prefer to keep a clear focus in the paper (instead of going towards a general review), we did not further elaborate on the potential reasons for SCE patterns in the treatment differing from the control. Instead, we refer to reviews about relevant mechanisms in the last paragraph of Section 4.3, which reads as follows:
The above list of potential mechanisms that can alter the moisture response of SCE when the precipitation regime changes is of course incomplete. Several other mechanisms can play at different levels (from community level to soil and microbial level). It is beyond the scope of this study to go into detail about all potential mechanisms. For reviews about various changes in the ecosystem under altered precipitation regimes, we refer to Borken and Matzner (2009), Schimel et al., (2007) and van der Molen et al. (2011). Here, we want to emphasize the need for a holistic approach in experiments that aim to elucidate how SCE is affected by changes in precipitation regime.
b) This has now been dealt with in the new conclusion section (conclusion section is provided in the answer to comment 37).

49. Technical comment
- Pag 872, line 7: The sentence “photosystesis or ecosystem respiration are can be measured at high frequency” need to be rewritten.

This was indeed a typing error. In the revised manuscript ‘are’ is removed from this sentence.

REFEREE 3

Initial Manuscript Evaluation Report The reviewed manuscript represents a fully comprehensive assessment about the suitability of extrapolating current relationships between soil respiration and environmental predictors (soil temperature and water content) in order to forecast soil respiration responses under scenarios of climate change. To test this idea, the authors make use of information from 38 precipitation manipulation field-experiments that cover a range of environments world-wide for which there are available data (some important climate-types, however, are poorly represented –tropical- or are lacking –boreal-). For each of these experiments, four different models are parameterized using respiration, temperature and moisture data from soils in control plots. Then,

temperature and moisture data from the corresponding manipulated plots are used as inputs, in the best control models, in order to predict soil respiration under an altered rainfall pattern. These predictions are afterward compared with actual soil respiration data from the manipulated plots. For the cases predictions didn't match observations, further analyses were performed to get insights on possible underlying mechanisms that would explain why the control model failed to predict soil respiration under altered precipitation; for this, different sources of variation among studies (climate, soil-type, frequency of observations: : :),as well as the time-course of the accuracy of predictions, are considered. The obtained results are highly relevant since they point out to the constraints on applying current relationships to predict soil respiration fluxes under altered precipitation regimes. In particular, the conclusions about the need to improve experimental designs (higher frequency of soil respiration measurements,more accurate assessment of soil water availability, and the consideration of both immediacy and legacy effects of climate extremes) will likely have a strong impact on future studies, and modeling approaches, to assess ecosystem responses to altered precipitation regimes. In summary, I consider the manuscript is of very high scientific significance. It certainly represents the first complete and spatially extensive test analyzing an important and open question related to the estimation of a key component of the carbon cycle under future climate conditions: to what extent, phenomenological models linking soil respiration to soil temperature and water content variability will remain unaltered beyond the current climatic window within which such relationships were constrained. On the other hand, the study focuses on the best currently available experimental system for testing such a question: an ample, worldwide collection of field-experiments involving manipulation of precipitation and mid- to long-term monitoring of soil respiration responses. And the test has been performed elegantly through a well-planned approach in which different modeling and statistical data analyses are applied, in a sequential scheme, to calibrate control models, to validate predictions, and to assess raised questions from the cases where predictions didn't pass the validation tests. Finally, the manuscript is concise and well structured, and the results are highlighted and discussed in an appropriate and balanced way, emphasizing the resulting perspectives and recommendations for future studies, and the identification of knowledge and approach gaps (e.g., the need that models account for the Birch effect, given the additional experimental evidences on the relative contributions of the heterotrophic and autotrophic components of soil respiration). —- As a minor change, I suggest the authors should clarify the (apparent?) contradiction between results that are differently highlighted in the Abstract or in the Conclusions sections. The Abstract stresses that "there are no serious problems associated with extrapolating current moisture responses to future climate conditions". This conclusionis based on the fact that the hypothesis was accepted in most (> 80%) of the cases studied (those for which models parameterized with data from the control plots -soil temperature and water content as predictor variables- did adequately predict soil respiration measured in manipulated plots). However, the Conclusions section underlines the striking correspondence between the cases for which the hypothesis was rejected and the cases with the highest frequency of soil respiration measurements. Consequently, the Conclusions highlight that "Our analysis demonstrated the limits to applying current soil moisture responses for predicting soil respiration under altered precipitation regimes". Is it that an inadequate parameterization of the models in the cases the hypothesis was accepted (because frequency of available data was too low) led to the wrong conclusion that extrapolation of current moisture responses to future climate conditions is suitable? Please, clarify this.

We thank the reviewer for this very positive report. The minor change requested has been dealt with in the revised version of the manuscript, in which the conclusion section was completely rewritten (see answer to comment 37).