Response to the interactive comment by Grace Pold

Cyrill Zosso and colleagues present an interesting look into how deep soil warming alters microbial biomass and high-level community abundance in a forest ecosystem. Overall, I think this is a very interesting paper and applaud the authors for their use of lipids rather than getting stuck in the mud with sequencing, as it provides a different perspective on warming impacts on microbial communities than is typically used. The paper is also well-written and flows well, and was enjoyable to read. There are a few areas I think the manuscript could benefit from clarification on. In particular, how and/or why certain comparisons for depth*warming interactions were made, and a better integration and justification for the brGDTs in hypothesis testing.

We thank Grace Pold for the constructive and insightful comments. Below we provide our point-by-point reply and we believe the comments will help us further improve the manuscript.

Main comments:

The scale on which the data are presented does not seem to match the scale on which the analyses were completed.

We were not quite sure how to understand this comment. However, cross-checking all figure units and calculations again, we did find that there was a mistake in the labelling of MBC graph and table. The values are reported in mmolC/gsoil, but should be nmolC/gsoil. For comparability, we double-checked which is the most widely used unit. As it is more often the case that MBC data are presented in ugC/gsoil, we would suggest to report them in these units. It is then easier to compare our values with other publications. Nothing changes in the trends reported.

It would be very helpful to post the R scripts with the manuscript. I found it hard to follow what was being included as a random effect (or whether random effects were nested) in the different results. Was a model of form "lipid ~warming *depth + (1|block/warm)" or "lipid ~ warming *depth + (1|block/depth)" fit? Or something else?

Below we post the model used and we will also publish the script alongside the manuscript (as well as the dataset) on ESS-DIVE repository. We did consider taking plots (n=6) nested within block, however we did not have replication within blocks. We did not include warm or depth nested within block as we considered them fixed effects.

"lipid ~ warming*depth, random=~1|block"

Related, there seem to be a lot of post-hoc tests, but it is not always clear how these posthoc tests were selected for completion. Why is the top 10cm sometimes compared to the bottom 10cm, and other times some intermediate depth compared to the deep?

We were not quite sure where the reviewer saw a comparison of top 10cm to bottom 10cm, but we think it could be the following misunderstanding. We think this could be related to e.g. lines 213-214. Here, we state the concentration in the top and in the lowest depth increment and a p-value. However, in this section we do not compare solely the top 10cm to bottom 10cm, but just mention these values as starting and end point. The p-value is from the LME model which was run through all depth points. If this was the problem, we suggest to reformulate these sections.

We only report post-hoc analysis including all depths for total PLFA concentrations (lines 226-228) and brGDGTs (lines 254-255), where we ran the function emmeans(). However, to simplify and reduce confusion, we propose to delete these two sections. These post-hoc tests are not discussed further in the discussion, since we generally focused on the results of the LME, which we explain in more detail below.

Or why is the cutoff for depth 30cm in some instances, and 50cm for others (ex. total PLFA vs. Actinobacteria)? Why is there only one p-value reported (ex. L229-230) when looking at all the depths below a certain point, and not one for each of the depths analyzed as show in the figure? Was the total lipid below a certain depth summed for each core to complete this analysis? [I think this would make sense from a statistics standpoint]

We generally report p-values from LMEs (except Post-hoc mentioned above). The models were first run for the whole profile. We then additionally ran the LME model on a subset of depths (either below 30cm for total PLFA or below 50cm for actinobacteria/brGDGTs) if the treatment*depth interaction was significant or upon visual inspection of the plots. Furthermore, we observed more pronounced treatment effects in the lower subsoils for several other variables as reported in other papers, such as the average chain length of solvent extractable lipids ('*Ofiti et al. 2021, SBB, Warming promotes loss of subsoil carbon through accelerated*

degradation of plant-derived organic matter') or d13C and free particulate organic matter (*'Soong et al. 2021, accepted for Science Advances, Five years of whole-soil warming led to loss of subsoil carbon stocks and increased CO2 efflux'*). That's why we think it is interesting that such effects were also observed in microbial parameters in the lower subsoil. Thus, the values reported in the manuscript are from the LME. When we sum up the values below a certain depth as suggested by the reviewer and run a t-test, the values are $p<0.01$, $t=-4.87$, $df=4$ for Actinobacteria and $p=0.054$, $t=2.7$, $df=4$ for brGDGTs (see table below). As this does not account for the blocks and depth, we suggest to continue reporting the values from the LME.

Second, please add more justification for why brGDGTs were measured (in terms of specific hypotheses) and how they should be interpreted. The authors mention that they turnover slowly…is this why they were chosen? If so, what does this mean for interpreting the data from a 4-year warming study? Since these are predominantly necromass, is the idea that microbes under heated and control conditions might be preferentially consuming these or warming might accelerate their turnover? Or is the idea to try and see if there is a signal in the brGDGTs that might indicate how microbial communities have overall changed in the past 4 years, which is not visible in the more rapidly cycling PLFAs? Or is the idea to capture predominantly the archaea community, which would not be captured by PLFAs?
Indeed, we will add some more justification for the use of the brGDGTs, especially in the introduction (line 73) and hypotheses. In the study by Ofiti et al. 2021 (SBB, Warming promotes loss of subsoil carbon through accelerated degradation of plant-derived organic matter) we observed that branched fatty acids were less abundant in warmed plots than control plots. Furthermore, the average chain length of fatty acids was lower in warmed as compared to control plots, especially below 55cm. These observations indicate that there might be decomposition or less input of microbial necromass. The brGDGTs were used as a more time-integrated and independent proxy to assess whether also these compounds, consisting of both microbial biomass, but mostly necromass, are less abundant in warmed plots. This would support the notion that there is either less input or more decomposition of microbial necromass. Thus, the idea is to have an additional proxy for the turnover/input of microbial necromass to soil.

Isoprenoid GDGTs (iGDGTs) are indicative of the archaeal communities. We did not report these values as iGDGTs were low abundant in our samples, for many samples below detection limit.

Minor comments:
Since the soil is warmed by a smaller degree in the top 20cm compared to below that, why not just discard the shallow soil data since it cannot fairly be compared with the deeper samples? Also, since lipids were extracted from less soil in shallow compared to deep samples, the deep samples could just be more representative of deep soil and therefore easier to detect a difference in.
We did observe a similar respiration response at all depths, indicating that the microbial response might be similar despite the difference in temperature magnitude. Nevertheless, we agree that 4°C warming likely affects the microbes and the community differently as compared to only 2.6°C. We do think that reporting all data gives added value, but we propose to discuss this caveat in more depth in paragraph starting on line 318.

We propose to adjust line 319: We did observe a similar respiration response at all depths (Hicks-Pries et al. 2017), indicating that the microbial response might be similar despite the difference in temperature magnitude. Nevertheless, temperature controls the reaction rates of microbial enzymes, which in turn can affect microbial abundance (Allison et al., 2010). Furthermore, incubation experiments show strong effects of temperature magnitude on soil respiration, including subsoil (Yann et al. 2017, Scientific Reports). Thus, the higher magnitude of warming below 20 cm in our experiment might be partially responsible for the observed difference in the microbial response between top- and subsoil.

As mentioned by the reviewer, the amount of sample extracted could mean that an effect is easier to be detected in subsoils. We think this is well worth mentioning as a limitation, however would also argue that by thoroughly homogenizing the sample this caveat should be minimized. Furthermore, we also observed more variation in topsoil in parameters where a constant weight was used for the analysis (e.g. carbon), indicating that the variation is likely naturally higher in these topsoils as compared to the subsoil.

L135: why add the standard in after lipid extraction, and not directly to the soil so that the authors could get a better idea of extraction efficiency for the different soil depths?

We agree that having a recovery standard can be helpful to have an idea of changing extraction efficiency with depth. However, during implementation of the method in our laboratory we tested having the PC19 standard as a recovery standard alongside the D39C20 standard. We decided to only use the D39C20 standard, because the reproducibility was considerably better when only using the D39C20 standard. Whereas the coefficient of variation was generally below 10% using D39C20, the coefficient was around 20% when using both, D39C20 and the PC19.

Could figure 4 be presented as an ordination instead? There is a lot to digest here.
We did use multivariate analysis for the exploration of the data, which did help to identify some of the trends (e.g. actinobacteria, anteiso/iso), but was rather confusing for the purpose of giving a good overview. However, we are happy to reconsider this figure and try and make it more approachable.

Grammar/style
The "gram" in Gram positive/negative should be capitalized, as it refers to someone's name (Hans Christian Gram)
We will adjust this accordingly throughout the manuscript.

L177: please mention what kind of post-hoc test was used.
As mentioned above, we propose to delete the two sections where depth wise Post-hoc tests were conducted and rather focus on the LME results.

L281: correlations are generally reported as R (for Pearson) or rho (for Spearman); R^2 is the coefficient of determination.
We will adjust this accordingly.

L300: there are also a lot of unknowns with respect to extraction efficiency of chloroform fumigation extraction. It almost certainly underestimates the C content of high surface area:volume small cells, as it predominantly captures cytoplasm.
We will be happy to look into this more and add this limitation with a relevant reference. If the reviewer has a good literature suggestion, this could be very helpful.

Please also make sure to report the F/T/Z statistic and degrees of freedom, preferably in the text, or otherwise in a supplementary table.
We will report this information accordingly.

Whenever someone says something increased/decreased with depth, it sounds like depth has been treated as a continuous rather than categorical variable. So please try to avoid this.
Erroneously, we did treat depth as continuous rather than categorical for the analysis. We ran the analysis with depth as categorical and will make the necessary changes. Mostly, the effect of this adaptation was minor, not affecting the interpretation of the results. But in some cases (e.g. overall concentrations of PLFA), the treatment effect did change, leading to minor changes in the manuscript. Find below an overview of the differences for some variables:

PLFA

	depth categorical		depth continuous	
treatment	0-90	30-90	0-90	30-90
p	0.045	0.005	0.057	0.002
f	4.334	9.668	3.820	12.011
df	1.000	1.000	1.000	1.000
depth				
p	<.0001	<.0001	<.0001	<.0001
f	41.950	10.213	283.552	62.123
df	8.000	5.000	1.000	1.000
interaction				
p	0.147	0.947	0.020	0.818
f	1.650	0.227	5.835	0.054
df	8.000	5.000	1.000	1.000

Actinobacteria relative

LME	depth categorical		depth continuous		t-test	depth summed up below 50cm
treatment	0-90	50-90	0-90	50-90	p	0.008
p	0.577	0.004	0.635	0.004	t	-4.870
f	0.318	11.578	0.228	10.814	df	4.000
df	1.000	1.000	1.000	1.000		
depth						
p	<.0001	0.033	<.0001	0.023		
f	7.587	3.861	24.175	6.171		
df	8.000	3.000	1.000	1.000		
interaction						
p	0.239	0.726	0.056	0.374		
f	1.383	0.443	3.846	0.833		
df	8.000	3.000	1.000	1.000		

brGDGT

	depth categorical		depth continuous		t-test	depth summed up below 50cm
treatment	0-90	50-90	0-90	50-90	p	0.054
p	0.278	0.004	0.422	0.027	t	2.701
f	1.216	12.128	0.655	0.027	df	4.000
df	1.000	1.000	1.000	1.000		
depth						
p	<.0001	0.001	0.000	0.395		
f	9.991	9.132	16.768	0.395		
df	8.000	3.000	1.000	1.000		
interaction						
p	0.206	0.866	0.093	0.617		
f	1.467	0.241	2.945	0.617		
df	8.000	3.000	1.000	1.000		