Reply to review of Simona Castaldi:

Review of the manuscript "Smallholder African farms in western Kenya have limited greenhouse gas fluxes" by D. E. Pelster et al.

General consideration The paper presents a nice dataset of soil GHG fluxes measured throughout dry and wet season in 59 locations in Kenya. There is no doubt that such measurements are surely missing and are necessary to better calibrate emission factors/ models of C and N cycle and GHG fluxes in tropical areas. What could have been interesting, but was not specifically analysed in the paper, would have been to have a value for background emissions (far from fertilization inputs) and a value of EFs for fertilization events. These might have been compared with IPCC, or other approaches which rely on the two different values (background and EFs). I understand that this paper could be seen as a first step into this direction, especially for what concerns average background fluxes. Less clear from the results is if a more intensive sampling approach is required to really provide reliable EFs and how representative are the presented data of the GHG emissions related to management practices. Maybe some comments on this can be added in the discussion.

Thank you for your review. We agree that it would have been interesting to have set up a study to examine the effects of fertilization events using comparative "control" and "treatment" plots, however, this has been done in a few studies (see Hickman et al. 2015. JGR: Biogeosciences. 120: 938-951). We have also submitted another paper that also does this (Rosenstock et al. submitted to Journal of Geophysical Research – Biogeosciences). However, what we believe is missing (and a logical first step) is a study that examines the current practices and what they mean for GHG emissions.

Regarding the last comment given above, obviously a more intensive sampling approach would provide more reliable EF for current management practices, however sampling 59 different farms (many of them with very difficult access) already required a massive, concerted effort. More frequent sampling simply was impractical. There are some other studies that mention how sampling frequency affects the accuracy of the estimated annual flux (Parkin 2008, JEQ. 37:1390-1395; and Barton et al. 2015. Scientific Reports. 5:15912). We could definitely add a bit more into our discussion that incorporates the effects of sampling frequency on the estimates.

Abstract: Comments

A1) It is not clear in the abstract and in the title if the GHG fluxes presented are a net ecosystem exchange or soil fluxes. It should be specified.

Any tall plants were excluded from the chambers, therefore what we present would be the soil fluxes.

A2) Would "pasture plots" be a more suitable definition than "Grazing plots"?

Since the animals actively grazed these plots, we believe it is more appropriate to call these plots "Grazing"

A3) Similarly a treed plot is not an appropriate land cover definition? Agroforestry? Open savannas (grass with some trees)? Orchards? . . .

These were plantations (either Grevillia or eucalyptus).

A4) page 15302 Lines 18-20: This statement sounds odd. You have just said that emissions are very low, basically these systems are low emitters of GHGs. And this in one fact. The other fact is that crops are not able to take advantage of fertilizer addition as some other factor is limiting. So independently from the global warming the second issue is that production is scarce. And clearly you don't improve it just by pumping on fertilizers. To increase the nutrient use efficiency is really an issue for food security rather than for global warming mitigation here. On the contrary in highly productive systems (polluting systems) which respond fast to fertilizer intensity the two issue are really strongly related and precision farming is a potential solution for GW mitigation.

We believe that the reason why the fertilizer did not cause increased emissions is because the soils are very depleted and the small amount of fertilizer added was utilized immediately by soil microbes or bound by the soils (and therefore not available for production of N2O). Therefore we still believe that intensification will lead to lower yield scaled emissions and also improve the local food security.

Introduction: general comments The Introduction is well presented and objectives are clear. Table 1 – It might be more interesting to compress the info on time length of measurements in one column (for example 1yr-wkly, or 1 wet season – bewkly) and add a column with infos on the agrosystem type analysed.
We agree that it would be a good idea to add in this information.
In Table 1 specify what does the "Flux rate" range, you report, represent. It is not clear. For examples if you have just one site what is the range for? Tot emissions for different crop cycles on the same site? What about when you have more sites? Just specify what are the numbers we are reading.
The ranges are for the different replicates and treatments in the study.

Materials and methods
Fig 1- the figure as it is doesn't help 1) to localize the study area precisely, 2) to imagine the distribution size of the study area in relation to the geographical location, as no reference is available for the reader in the gray figure except the longitude. I suggest to zoom in the first figure of Kenya to show in which district/town area the star falls (we assume the reader knows Kenya is in Africa), and in the second figure it could be good to have the dots on a google earth kind of background with some reference points clearly shown to help other researchers to immediately identify your study area.
We agree that this map could (and should) be improved.
Comments: MM1) 15305 lines 1-2 – what you mean by "to be broadly representative of demographics and agro-ecological characteristics of other East African tropical highlands"? demographically speaking? Same average population density?
Similar population density and income (i.e. lots of poor smallholder farms).
MM2) 15305 line 3 – Could you specify in which "climatic zone" are the sites (adding the adequate reference)? It helps when categorizations are done in scaling up studies.
We can add in the Köppen classification to clarify what is meant by "climatic zone".
MM3) 15305 line 16. . .. When you define the soils with a specific classification name, specify which classification system are you using. . .USDA? Other? Cite the reference.
This was FAO system
MM4) Table 2 – In the main text some clarification and better explanation for the brief description given for the 5 land classes is needed. What is moderate size for you? 1 hectar? 10 hectars? What are degradation signs?... How slopy is the slope? Would that contribute to have erosion?...just to understand what are we looking at. You need to specify in the legend the soil depth of the analyses you present in Table 2 and the time frame of the presented data? You sampled before starting? I assume C content is total C by CNS? Any calcium carbonate which might give you Tot C > org C? please specify in the column heading if it is total C or organic C.
Small farms were around 1-2 ha, moderate, up to about 10 ha. The degradation signs that we are talking about are signs of erosion that were visible on the MODIS images. Regarding the C content, it was total C, however with the soil types, pH etc, it would be very unlikely to include any CaCO3.
MM5) page 15307 lines 1-5. I generally do not like to read a paper where the key methodology necessary to understand the meaning of the results requires reading another paper. I think the authors should make an effort to summarize in a comprehensive and transparent way the criteria they are using to distinguish the land classes they will discuss later on and how these sum up to create field types and land classes. Maybe you can add some additional tables where we can see

the single parameters and the score they have for each category used to build up the discussed land/field types.

If we add in too much information, we fear that the methods section will become much too long. We had thought that we had reached a nice balance between providing enough details to understand what we did, while providing the references if the reader wants more details. We can add a few more details however if you feel that we were a little too brief.

MM6) 2.2 soil core incubation: It is not completely clear to me the procedure you used here, maybe you can explain better at the beginning what are you exactly aiming to before describing the procedure. Drying completely the soil and rewetting it creates a sort of extreme situation where a significant part of the N used by the system can come from the dead organic matter dry-wetting cycle itself. The flux of gas can fade away in a day or more. From the way you describe it you add water, close the jar and measure the efflux at 0, 15, 30 and 45 min. In my experience that flux is not representative of the baseline flux of a site. It is representative of post rain flushes. I understand that in order to increase the WHC you need to start from low WHC, but how do you use the number there after? Are they representative of which soil characteristic or potential, independently from the KNO3 addition?

We were trying to measure the potential for GHG emissions, so we purposely created an "extreme situation" and then measured the emissions. Also, we are not using this to represent baseline flux of a site, but rather just to compare the potential amongst the various sites.

MM7) 2.3 Field soil GHG flux survey It is not specified the number of chamber replicates you use for each plot. The only time a number is mentioned is when you specified that you pooled gas samples from 4 chambers in one syringe. Does this mean that you had 4 chambers x each site? If you pool the gas at each sampling time in one sample it means that basically you are measuring just one gas sample per plot? No replicates whatsoever? Spatial or experimental (lab replicates of the same gas sample)?

The reviewer is correct that we had no "replicates" (only pseudo-replicates) from each site. We examined the trade-off between having spatial replication at each plot versus examining many more plots and decided to pool the samples from the individual plots, but do more plots

MM8) It doesn't seem that the gas sampling pattern follows any specific management practice timetable. Could clarify the rational for this. To clarify what I mean, we know that in particular for N2O, but also for CO2, fluxes of gas occur when something happens (manure or mineral N addition, tillage, crop collection). The flush lasts for a time which can go from few days to some weeks and is proportional to the magnitude of the management practice and soil characteristics. It often makes most of the annual total GHG flux. So to miss the flush means to underestimate the overall crop flux. Isn't this something to take into account when rescaling the magnitude of fluxes in your system? It could be important to have some clarification in the procedure on the relative importance(or irrelevance) of this issue for your system.

We sampled weekly, regardless of management practice. As you mention, this results in some the limitations (e.g. potential for missing peaks, and how this may lead to underestimates). We can add some clarification on these points.

MM9) 2.6 Environmental data It is not clear for soil moisture and temp how many probes you used? Were they fixed near the climatic stations?

We used not only the two weather stations, but also measured soil temperature and moisture at each site, at the time of gas sampling using a ProCheck handheld datalogger with a GS3 sensor (Decagon Devices). So when we compare the emissions to variables like soil moisture, we use the data from each individual plot, but when we provide rainfall data, it is from the weather stations.

MM10) 2.7 Plant production It is not clear what are you doing here. Why are you sampling only 9 plots? Why not 59? What are this only 9 plots for? What are they representative of?

We did not have the capacity to measure all of the sites and so we decided to focus on just the annual crops. We also tried to sample the most common annual crops (i.e. maize and sorghum)

MM11) 15311 line 14. Better use "field"experiment rather than "in vivo", this latter expression is used for biological rather than biogeochemical experiments.

This is a good point and easy to change.

Results: comments

R1) You are making a statistical comparison among land classes. I don't remember I have seen anywhere specified the number of sites falling in each of the land classes. I assume it is an unbalance statistical design. How much unbalanced? Are some of the classes over represented?

The N for the different land classes are as follows: landclasses: 1) lowland subsistence farms with degradation signs (N = 7); 2) lower slopes, moderate sized mixed farms (N = 8); 3) mid slopes, moderate sized, primarily grazing / shrubland (N = 10); 4) upper slopes / highland plateau, mixed farms (N = 22); and 5) mid slopes, moderate sized mixed farms (N = 12).

For field type: field type 1 (N = 17), those with a low score were considered field type 3 (N = 19) and those intermediate plots were assigned a field type 2 (N = 23).

And for vegetation type: treed/bush (generally plantations of either Grevillia spp or Eucalyptus spp) (N = 7), perennial grasses/grazing (N = 15) and annual cropping (N = 37).

R2) page 15312 "there were no detectable differences in N2O or CH4 fluxes between crop types" are you considering in this case differences in crop types within each class or independently from the classes?

These were done independently from the samples because the imbalanced design caused too many of the combinations to have too low N values.

R3) If I understand correctly the field type 1, 2 or 3 combines all the classification scores. Correct? Is it the case that some of the classification scores which build the same field type go into opposite directions in terms of their impact on N2O fluxes?

This is not likely. Field type 1 would be the most highly managed (so highest inputs, both fertilizer and carbon), and were found to be the most fertile (see Tittonell et al. 2013 – reference provided in the paper for more information).

R4) page 15314 lines 16-19. It would be interested to understand if considering the single sites, the management effect would still be not significant on N2O fluxes, which seem to double in the wet season compared to unmanaged sites.

This is a good point. It is possible that if we used a single site and compared the emissions for fertilized and unfertilized plots, we may see significant differences. As we mentioned earlier, this was done in another study that we did (Rosenstock et al. submitted to Journal of Geophysical Research – Biogeosciences; also Hickman et al. 2015. Journal of Geophysical Research – Biogeosciences), both of which found either no effect or a very small effect of management.

Discussion

D1) 15320 – lines 6-12. Given the very low emissions from these soils, would such a system (cores) be necessary to define management practices, beyond the general criteria used to predict high/low N2O emission potential of agro-sites? (drainage class, C content, fresh C inputs, structure and bulk density, average water content from rainfall or irrigation. . .the usual stuff used in other continents to reason on N2O emissions vs management). Beside, despite the correlation, I assume we cannot predict emissions in the field from emissions from cores, can we?

As noted above, we cannot use the incubations to predict field emissions (which we can state more clearly in the discussion). We do think though, that this can still be used to confirm (in a comparative study only), which sites have a higher potential to be hotspots (and yes, it would probably be related to many physical/chemical properties such as BD, water content, C content etc).

D2) page 15321 lines 1-9. I think that the authors should discuss how much influence might have the sampling design on the observed "lack of difference" of GHG emissions among land/field types. GHG emissions and in particular N2O emissions are very spatially and temporally variable. Moreover, in agricultural ecosystem, the budget is strongly linked to any form of N input to the system, with emission peaks following N inputs and requiring intensive analysis after fertilization to avoid missing them. Could the sampling design (time, replicates) have been insufficient to have a complete picture of peak events? Can you discuss this, it is important, it is the drama of each study in agrofields no matter the geographical area. Also, the way the analyses are presented tends to average the fluxes within class blocks derived from your classification system, which includes many parameters a part from fertilization. What happens if we consider only fertilization intensity vs fluxes? Could the sampling design contribute to flatten the results also in this sense?

This is a very good point and should be examined more thoroughly. As we mention above, there are two previous studies that examine the effect of sampling frequency on the accuracy of the emissions estimates. These both show that is systems with high variability, the estimates from weekly sampling can be off by quite a lot. This is an important point and should be added to our manuscript. However, the most highly variable GHG in terms of temporal variability is likely N2O, which had very low cumulative fluxes for almost all of the plots. So even though we may be off quite a bit (in terms of percentages), this would translate to a small “absolute” value.