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

Anonymous Reviewer:

The paper presents a set of flux measurements measured using traditional chamber methods from an area not well accounted for in current GHG measurement budgets (rural Africa). It is true that data is required from these areas to better account for gas fluxes in the region; however the broad approach and limited number of measurements used in this paper to estimate annual flux values is unlikely to represent well the complexity of the processes occurring at such a large scale. I believe that several of the methods in the paper are flawed and interpolation of the data points is too simplistic to provide reliable tier two annual estimates without large uncertainties, which are missing throughout the study. I would encourage the authors to defend their methodologies and improve upon their uncertainty estimates before publication. Although I have several concerns with the methodology used, the data set may still prove valuable as a starting point to those investigating fluxes in this region. The paper is generally well written with a few grammatical errors creeping in. If the authors can justify their methods and improve on describing uncertainties in their measurements and interpolation of data, then I would consider the manuscript worthy of publishing.

Comments:
Introduction
15302 L26 A reference to IPCC values/estimates may be helpful
We already reference the Vermeulen paper (2012), which provides a very robust estimate of the contribution of agriculture to global and continental emissions.
P15303 L3 replace 2/3rds with 66 %. All other numbers in the text refer to %
Done
Smallholder farms may account for 80 % land coverage, but do they account for 80 % fertiliser use?
Likely not, but our objective was to focus on smallholder farms, which cover a large area, but are rarely investigated for their role in GHG emissions.
What is the difference between industrial/subsistence farming inputs?
In N2O studies the N (kg ha-1) content of the fertiliser is very important. Without this it is impossible to tell the EF% of the fertiliser events for N2O. 20 Kg ha-1 of manure will have a very small nitrogen content compared to typical European fertiliser applications.
This is very true, however we wanted the farmers to use their normal practices, which typically is quite reactive, meaning that we usually learned about management practices a few days to a week after they occurred. Also, we did not have the manpower to increase sampling to a higher frequency.
Table 1 Units should be consistent in the table.
The reason for different units is that in some cases there was insufficient information to calculate annual fluxes, in which case we used the cumulative emissions for the period, or the mean rate, depending on what was given in the reference. We could convert the mean rates or cumulative periodic fluxes to similar units, but converting them to annual fluxes would be incorrect and result in far too much uncertainty, while converting the annual fluxes to a mean would result in too much lost information.
There is no mention of the large amount of agricultural goods produced industrially in Kenya. Tea, coffee, cabbages, onions, mangoes etc. . . Although these areas are not covered in the study, the introduction describes Kenya as a sub-Saharan area with low fertiliser input and low productivity. This is not the case across the entire country.

While this is true, our objectives were to measure emissions from smallholders. Also, according to government of Kenya data, in 2009 the specialized agricultural goods accounted for less than 25% of fertilizer use.

Methods

Incubation study

The soil cores were air dried for 2 days at 30 degrees Celsius. Is this to replicate realistic environmental drying conditions or is this beyond what can be considered natural?

Removing the cores from the soil and then drying and placing them in jars would change the oxygen content and aeration of the soil significantly (among other physical and biological changes). This would have a serious impact on the gas producing microbiological processes which may nullify the validity of any results obtained using this method. If water was added to the samples in a single step, this does not reflect a rainfall event well and the fluxes measured immediately after will not either. If the soil incubation work is to be included in the article, the method and any assumptions made would need to be better described.

We now state that the incubation study was to compare "potential fluxes", which may not be indicative of the flux rates in the field (see response to reviewer 1)

Field study

Pooling separate air samples from chambers can only reduce representation of spatial variability from the plots. With individual chamber measurements the spatial variability can be assessed statistically (or at least attempted). If GC sampling costs or time were an issue, then decreasing the sample number may have been acceptable; however, in pooling the samples any information on linearity of regression in the individual chambers or spatial variability within the plot is completely lost. Rather than deal with several large sources of uncertainty in scaling the data (spatial variability and regression), using this methodology the uncertainty in the sampling methodology becomes completely incalculable and ignored. Although this method has been published and peer reviewed it is my opinion that it cannot be used to scale fluxes spatially due to large unaccountable uncertainties and possible statistical bias. This method prevents the propagation of any kind of uncertainty in spatial interpolation when calculating cumulative fluxes.

We agree that the pooling results in a loss of on-site variability, however this is an accepted method and is similar to the principle used when compositing soil samples for analysis (which is also generally accepted). It is therefore the opinion of the authors that this method can be used to scale fluxes.

The assumption that a loss of CO2 represents leaks in the chamber may work in general, but the pooling of samples and the inability to determine which chambers leak, or if the different gases behave differently in each chamber is a real weakness of the method.

It is not perfect, but as shown in the paper by Arias-Navarro it still provides estimates within +-4% of what is found when sampling the chambers individually. We have added in the accuracy estimates to the M&M

Throwing out data with lower values than the precision of the instrument is not good practice. What is the instrumental detection limits for each gas? Is this consistent on a day to day basis or does it change? How is it calculated?

We did not throw out data with values lower than the precision of the instrument, rather we assumed the flux was 0. In terms of the calculation of the cumulative fluxes, the effect of this will be negligible. The cases where we threw out data was where there was a poor fit, indicating contamination of the sample or leaky chambers or some other sampling error.

All data should be included and thus the instrumental precision is then un-biased as it is equally positive and negative. In literature CH4 uptake is generally believed and N2O uptake is not. In this paper all N2O fluxes below zero seemed to have been removed while CH4 is allowed (perhaps because it is generally accepted in literature?). This can bias the results. In Figure 2 I see nothing in the CH4 measurements that can prove anything other than instrumental noise is being

measured. For N2O it is impossible to tell as the axis has been cut off at zero. Were no negative fluxes of N2O recorded during the study?
The reviewer is correct in that we allowed for uptake of CH4, but not of N2O or CO2 because this is what is generally accepted in the literature. We can explicitly state this decision process in the methods. Also, the reviewer is correct in noticing that the CH4 flux in Fig 2 is likely just instrumental noise and we can also clarify this as well.
How can detailed regression analysis be used if the chamber samples are already pooled? Each chamber has its own linearity for each gas type depending on conditions.
This is an accepted method and one that has been shown to provide very similar estimates to the traditional method of calculating a mean of non-pooled samples for all of the gases we measured. The regression is done on the headspace concentration as a function of time. Even in a single chamber there will likely be high spatial variability in terms of sources/sinks for some of the gases (i.e. N2O) and therefore the headspace of a single chamber is already an "average" of the flux rates from the various sources.
How was temperature change accounted for within the chamber? 45 min long chamber times may result in very large temperature changes during the enclosure times, especially during hot days. This can change the physics within the chamber in a way that can affect flux calculations. (Air density, expansion of air, large pressure changes, etc..)
We measured chamber temperature at the start and end of the 45 min deployment and used the mean to correct the concentration estimates. We also used insulated chambers covered with reflective tape to minimize temperature changes. Although this is not perfect, it should account for much of the changes within the chamber listed above.
No uncertainties were calculated for the entire flux process. This is a real weakness in the method which should be addressed if the data is to be published. Any uncertainty in scaling up at the plot scale is lost in the pooling sample method and no estimate of temporal uncertainty in linear interpolation of the measurements is even discussed.
We have provided some uncertainties in the paper (i.e. mean + SEM provided in the tables). However, as the reviewer noted, we lack the information to provide information of variability and hence uncertainty at the plot level. However, we were not interested in variability at the plot level, but rather the variability between sites (and then landuses / landclass etc). The comment about temporal uncertainty however can be added as there are several studies that examine how sampling frequency affects the accuracy of cumulative flux emissions (Parkin 2008 and Barton et al. 2015). We know that weekly sampling may miss peaks, particularly for N2O immediately following fertilization, but this can be added to the discussion.
At the end, all of the plots are averaged out to give a range of uncertainty, but each of these numbers should also have a very large uncertainty associated with it. This number should then propagate through. From the results presented in this paper it is impossible to tell how precise the study has been in its estimates of annual fluxes. Is the method even fit for purpose?
We used published methods. If the reviewer questions the method, perhaps the reviewer can submit a critique of these methods. As it stands however, we believe that pooling the headspace samples of 4 chambers is no different from sampling from a single chamber that is 4 times as large, only we have better spatial distribution.
Results
Figure 3 No error bars are included in any of the measurements.
We felt that the figure was already "cluttered" and that the addition of error bars would only confuse things further and so we left them out. We still believe that it is the correct decision.
Table 2: should Bulk density have units of mass per volume?
Yes, the units provided here are incorrect and should be changed.
More information is required on what the CO2 measurements are actually measuring.
See response to reviewer 1, we are measuring primarily soil respiration as in most of the chambers (grazing plots aside) excluded plants. In the grazing plots, the grass was typically

grazed so heavily that they were less than 2 cm high, so we were again likely measuring primarily soil respiration.

Do they contain some plants or just soil? Why measure CO2 from the chambers? What do these measurements tell you?

See above.

The weekly measurements are likely miss peaks in N2O emissions from fertilizer events which can last less than a few days. Any attempt to do a cumulative annual budget for N2O emissions should do more regular measurements at least around fertilization dates.

We agree that this is a potential issue with our methods that needs to be expanded. As stated above, there are previous papers that have investigated how sampling frequency affects the accuracy of annual cumulative flux estimates.

Diurnal effects are not observed due to the manual chamber method being used during the day. Nocturnal emissions will have different temperatures (and light for CO2) which may affect processes in the soil. Differences in night/day fluxes in Africa may differ from those observed in the more commonly measured areas (Europe etc. . .).

Diurnal patterns in emissions were definitely not observed as they require either automatic chambers (these would require electrical power at the farms, which generally are too remote to have any electricity) or they require people to deploy the chambers multiple times per day/night. This was deemed to be unsafe for our staff and therefore, any investigation into diurnal patterns was considered to be beyond the scope of this study. However this could be investigated in subsequent studies.

Figure 4 Explain what the box plots represent. Quartiles of 59 points?

The boxplot is a fairly standard plot with the mid-line showing the median, the boxes showing the first and third quartile and the "whiskers" extending to the maximum (or minimum) value within the upper (and lower) fence – which is defined as: 1.58 * IQR / sqrt(n). This is roughly equivalent to 95% confidence intervals. We would prefer not to have to explain all of this in the figure caption as we believe it is unnecessary.

Discussion

15321 L15 Assuming that the measurements scale to a continental scale is highly optimistic.

We agree that this is optimistic, which is why we state that it might be valid at that scale. We should however, clarify that this is for smallholder farmers only.