I raised several issues in my previous review, and thank the authors for their efforts at addressing them. Overall, I think the discussion is much improved. Unfortunately, I believe my concern regarding the experimental design needs to be further addressed. I do believe that revision is possible, and although it will likely require new statistical analyses, it shouldn't be overly difficult once the appropriate design is identified.

Below I list four general comments, of which the first two are most important, followed by responses to some of the authors' responses to my earlier comments, followed by specific comments. Apologies again for the rather verbose review.

GENERAL COMMENTS

1. One of my main concerns was regarding the robustness of the conclusions that can be drawn from this experimental design (Author response 17). The authors have removed the seasonality analysis, which I appreciate, but I believe that important issues remain. Some of the problem is in the language used to present and discuss results rather than in the underlying science, but science issues remain.

- This remains a study using only one year of data, though I don't think this is in itself a problem, it is a limitation. The authors argue in their response 18 that a previous 14-month study (Petersen et al. 2012) should also be considered, and that study does provide context for the results presented here (though note that any results from Petersen et al 2012 that are important for understanding the rationale for this study or for understanding its results should be summarized in this manuscript). However, Petersen 2012 is not part of the analysis, and the current manuscript doesn't provide any insight into the effects of interannual variability (also it seems that this was perhaps an abnormally warm and wet year). That's not a problem--I think the authors have done a good job of removing generalizing statements about seasonality from the revised manuscript--but I do think that a statement noting that the analyses of spring and fall are based on data from a single year should be included in the conclusion as part of an acknowledgment of potential limitations of the study.
- This remains a study that lacks traditional replication of treatments, and I think this issue requires substantial additional revision.
    - The authors argue that RG1 and AR1 should be considered as independent units because they differ in land use history, and were treated as such in Petersen et al. 2012. I'm afraid I'm not convinced—the design is an illustration of the traditional definition of a split plot: two adjacent fields with different treatments (or, as in this case, treatment histories), but a shared geographic location, and thus conditions that have been formed by a history of shared state factors (sensu Hans Jenny). As such they are simply not independent sites. The map in figure 1a, which presents AR1/RG1 as a single site, illustrates the point rather well, and as I understand, sampling positions in the two fields were only 10-20m from one another. One of

the clearest issues here is the nearly identical temporal patterns of the water table depth—the kind of shared environmental variable that split plot analyses were created to accommodate. I think it's fine for the authors to describe patterns at each of the RG and AR locations, but I think the experimental design makes any formal comparison between land uses impossible. If the authors want to examine site differences, I think it would have to done using an n of 3, treating RG1/AR1 as a single site with higher within-site replication, but it would be a good idea to consult a statistician—there may be other options. Because of this issue, I think the graphical analysis may also need to be redone—perhaps organized around sites rather than land uses—if I am correct in my understanding that independence is an underlying assumption (as I mentioned in my previous review, I am not an expert in graphical analysis). In any event, I recommend consulting with a statistician before proceeding with additional revisions.

- A related note: the very big site differences between RG2 and AR2 (e.g., in carbon stocks) highlight the rationale for using split plot approaches (such as the one the authors use for their fertilizer treatments): such approaches avoid treatment or land use history differences being confounded with site differences.
- The authors also argue that differences between land use was evident without any statistical support, which is a fair argument. I think it's fine to discuss possible effects of land use qualitatively, but important to include and clearly highlight the caveat that the study lacks independent replicates for the land use treatments.
- I don't think these issues mean that the graphical model cannot be applied or provide novel results, but the model needs to be applied using the appropriate experimental design, and can't treat RG1 and AR1 as independent sites.
- In Response 17, the authors note that land use was not analyzed statistically. However, the generalized linear mixed model includes land use ("crop") in the interaction term used as the fixed effect (line 248-249), which is inappropriate given the replication concerns regarding RG1 and AR1.

2. This leads me to the language presenting results. The qualitative results need to be more clearly communicated as such—the way many results are presented here implies that there has been a quantitative statistical analysis conducted. A very simple example is line 328: "Mineral N concentrations were greater at AR1 compared to AR2". In the scientific literature, such a statement implies a significant statistical difference in mineral N concentrations between the two sites. If a statistical test was conducted, a P value should be included (in general, there should be much more inclusion of p values in the text). If it wasn't, small changes to the language need to be made that make that clear, e.g., "Mineral N concentrations appeared to be greater at AR1 compared to AR2". Another small example might be line 334-35: "There was variation at depth in the soil, which could not be explained by fertilization." The phrase "could not be explained" implies that a statistical test was attempted, but was not significant. If, instead, this is qualitative interpretation of data, it should include more qualifying language, e.g., "There was variation at depth in the, which did not appear to be related to fertilization."

In fact, there are numerous instances of this kind of presentation of results throughout the manuscript; here are just a few examples (there are many more involving site comparisons, temporal comparisons within sites, fertilizer effects, etc., and all should be addressed):

> L316: "There were only minor differences. . . between seasons" implies that seasonal differences were tested; can be changed to something like "If there were any differences, they were likely minor. . ."
>
> L 326: "The residence time for mineral N in the soil solution was generally longer at AR compared to the RG sites" implies that site differences were tested; can be changed to "The residence time for mineral N in the soil solution generally appeared to be longer at AR compared to the RG sites"
>
> L328: "Mineral N concentrations were greater at AR1 compared to AR2" implies that site differences were tested; can be changed to "It appeared that mineral N concentrations may have been greater at AR1 than AR2"
>
> L 329: "Fertilisation increased NH4+-N and NO3- -N concentrations" implies that differences between fertilized and unfertilized split plots were tested. Can be changed to "It appeared as though NH4+-N and NO3- -N concentrations increased following fertilization"

One quick way around this issue would be to state in the methods section that all interpretations of results are qualitative unless accompanied by a p value, and then to include p values where statistical tests were conducted (in any event, I encourage the authors to include P values for all statements reflecting the results of a statistical test). But a clearer solution would be 1) to test differences statistically if they have not been and include p values, and 2) to change the language describing results to make it qualitative, using words and phrases such as "may," "appears to," "might," etc. to describe apparent patterns that cannot be tested or aren't important enough to test. I don't think it's necessarily a problem to present results qualitatively (though for any results that are important to the central hypotheses being tested, it's always good to include statistical analyses), but it needs to be very clear when differences were tested and when they were not.

3. I think there could be more detail in the discussion of figures 3-7, which represent a lot of the data presented in this manuscript. Briefly, the text often invokes a relationship between the water table and a) N2O concentrations in the soil or b) N2O fluxes at the surface, but the relationships described in the text are not obvious to me from examining the figures—the spatio-temporal relationship between the water table depth and N2O concentrations seems quite different in different sites and seasons. Additional text walking the reader through the authors' interpretation of these figures would be helpful. I made related specific comments in my last review, and detail some points below in my response to the authors' response #46, as well as in specific comments for lines 494, 501, and thereabouts.

4. It seems to me that the two conclusions that can be clearly reached from this study are the rejection of the hypothesis that that FeS_2 oxidation coupled with NO_3^- reduction was an important driver of N_2O emissions, and the identification of the capillary fringe as an important predictor of surface N_2O fluxes. If the authors agree, the conclusion should be revised; as it currently stands, the second of these findings is not discussed at all in the conclusion, and the description of the first conclusion gives the reader the impression that the hypothesized mechanism is possibly and maybe even probably not trivial.

Minor comment: the response to reviewers letter was sloppy – a number (maybe all?) of the line numbers of revised text referred to in the response to reviewers are incorrect, there are typos in the quoted text revisions (e.g., response 50), and sometimes there's no indication of what revisions were made or where they can be found in the text (e.g., response 49). I emphathize with these kinds of errors, especially after getting a fresh batch of comments from a new reviewer late in the publication process, but anything you can do to make the job of the reviewers easier is appreciated.

Responses to specific author responses

Response 20: I think the authors missed my point here, which was that the manuscript seems to only conduct a qualitative analysis of the relationship between changes in water table depth and changes in surface N2O flux, rather than including a statistical analysis of the relationship between the two variables that could provide quantitative insight into the importance of variation in water table depth for N2O surface fluxes. I agree that the manuscript does provide a graphical analysis showing that capillary fringe N2O is the only significant predictor of N2O surface fluxes at several sites, but being statistically significant and being important are not necessarily the same thing, and the graphical analysis does not include any analysis of the relationship between water table depth per se and N2O. I don't think that the authors need to do what I am suggesting here for the manuscript to be publishable, but it seems like a missed opportunity.

Response 21: Thank you for the response, though part of my concern was not addressed. This concern is largely related to my general comment about the language used to describe results. My concern here had been that there were multiple instances in the manuscript (I cited them in the original comment as including lines 380 and lines 411-412) where temporal variation in surface N2O flux at a site was attributed to fertilizer effects, or where fertilizer effects were excluded as a cause of variation in surface N2O flux. However, the relationship between temporal variability in surface N2O fluxes and fertilizer applications appears not to have been tested statistically.

Response 25: The authors argue that "The graphical model results (Figure 7) did show increasing N2O emissions with declining, as well as increasing WT depth that depended on soil N status." I do not think that the graphical analysis included a water table depth variable?

A separate point: I also don't understand what is intended by "increasing WT depth that depended on soil N status" but if the point is that N2O_WT depends on soil N status, that would be an interesting result that deserves more discussion. If instead it's referring to AR-autumn, where the N2O flux is determined in part by soil N status, then ignore this second comment of mine.

Response 39: I'm glad this was caught. I'd encourage the authors to carefully review the script for producing all figures again if they haven't already.

Response 46: I wonder whether this explanation may also hold for DOY 252 and 259: it looks to me as though there are elevated N2O concentrations in surface soils, and no clear connection to the capillary fringe. Wouldn't that suggest a topsoil source for all of the highest surface fluxes at AR1 during Autumn—a finding supported by the graphical analysis? The text implies that the high WT is responsible for the high emissions for DOY 252 and 259 ("The subsequent decline in N2O emissions at AR sites coincided with WT withdrawal"), but I would think that anoxia in topsoil related to elevated precipitation during this period could be a more likely explanation. The variability in topsoil N2O concentrations looks fairly physically separated from the water table dynamics, and if capillary fringe is not a significant predictor for surface N2O fluxes at this site during autumn, why invoke a relationship between WT withdrawal and surface N2O emission declines?

I have further questions about the interpretation of the subsoil N2O concentration dynamics discussed in my comment on line 494 below.

Response 54: I think it's a good move to consider N2O as a time-integrated measure. My previous concern here was very much regarding the limitations of using point measurements of NH4+ and NO3- to infer N transformation rates. I think the framing of the revised discussion addresses my concerns.

Response 56: So the data presented in figure 2 are means (and standard errors) of intact cores containing either fertilized and unfertilized soils in RG1, but for AR1, it's a mean of intact cores containing only unfertilized soils? I don't understand the rationale for presenting the data this way (instead of, for example, presenting fertilized and unfertilized soils separately where appropriate). At the very least, the caption should clearly detail this odd fertilization treatment situation.

Response 57: It's your decision, but there isn't much fine-scale temporal variation in these figures; I'd think it's worth graphing up the full year and seeing what it looks like.

Response 59: Again, it's your decision, but would think it's worth including in SI as figures as well as data tables—it's just very difficult to read patterns in a data table. But thank you for including the raw data for researchers who may be interested in using it for their analyses.

New specific comments:

Line 23: perhaps define capillary fringe here: "…in the capillary fringe— [definition here]—was. . . "

Line 62: it would be helpful to readers if you could provide an explicit definition of capillary fringe here, given how important it is to the manuscript.

Line 105: what ridges? Probably need to provide more agronomic details

Line 124: Perhaps this should be "method of fertilizer application"? (as N fertilization was noted as an exception in the next sentence)

Line 127: I believe the erroneous NS fertilizer application in RG2 was made on the same date as the second slurry application; if correct I would clarify that fact by starting this sentence with "On the same date as the second slurry . .. " or "Immediately after the second slurry. . . " If the NS application was made on a different date, indicate the date.

Line 137: here it would be helpful to explain the choices of spring and fall as emerging from the patterns observed in Petersen 2012, and in the introduction briefly describe those patterns. It could be just 1-2 sentences in the introduction, and an introductory clause to the first sentence here.

Line 144: it would give a more complete picture if the dates of these exceptions were enumerated here.

Line 165: revise to "…at -20C until analysis, described in section 2.4.5."

Line 170-1: I'm afraid this sentence is hard to follow. It reads as though the diluted soil gas was first transferred to the exetainer, and then from the exetainer to the glass syringe. Perhaps also specify whether exetainers were filled with 10ml gas as they were in the chamber measurements.

Line 174: As a general rule I'd argue it's best to explicitly include all applicable methodological detail, rather than referring readers to an additional manuscript (within reason). In this case, my institution does not have a subscription to the European Journal of Soil Science, so my library had to obtain 24-hour access to Petersen 2014 so I could see the dilution calculations; I would include them in the manuscript.

Line 240-245: it would be helpful to specify if/how field temperature and pressure corrections were made to obtain surface flux estimates.

Line 297: It should be noted here that these characterizations were conducted after the fertilizer treatment of RG1

Line 380: Instead of "independent of fertilization" which suggests a formal analysis, how about "this variation appeared to be broadly similar between fertilized and unfertilized subplots"

Line 394 and following: this is one of many examples related to my general comment on presenting qualitative results, but the generalized linear mixed model appears to include only time, and not fertilizer, as a factor in the fixed effect, so statements regarding the effects of fertilization should be made qualitatively (e.g., rather than "with no effect of fertilizer amendment," use something like "and did not appear to vary in response to fertilization." Similarly, "no effect of N fertilisation was observed. Hence, the higher emissions were associated with site differences other than fertilization" could be changed to "fertilisation did not appear to influence N2O fluxes. Hence, the higher emissions were likely associated with site differences other than fertilization"). The paragraph starting with line 406 does a good job of qualitatively describing results.

Line 467: change to "In accordance with this effect of rewetting…" (in general, it's good to give "this" an object to make it clear what is being referred to)

Line 490: here and maybe elsewhere there are still instances where time of year are referred to without DOY (here parenthetical DOYs would be particularly helpful since its referring to figures using DOY on the x-axis). In addition, please specify that this accumulation is in the fertilized plots of RG2.

Line 491: change "significant" to "substantial" if the meaning is "a lot." Avoid using "significant" except when alluding to the result of a statistical test.

Line 492: if the contrast is that in RG2 there was only accumulation at certain times and depths but accumulation was everywhere and all the time in the AR sites, change "..accumulation of N2O in the soil" to "…accumulation of N2O across all soil depths throughout the spring" or something similar. Maybe include a reference to Figure 4.

*Line 494: as was the case with the previous version of the manuscript (see my original comment that elicited the authors' response 46 and my general comment 3), I'm not sure I see the pattern so clearly. The location of elevated N2O above 40cm in AR2 is attributed to the water table being higher than in AR1, where elevated N2O tends to be in deeper soils, and the authors argue that this points to a capillary fringe source for N2O. But in RG2 the water table is higher than it is in RG1 during the spring, but the N2O distributions are reversed: they are higher at depth in in RG2, where the water table is higher, and higher in shallower soils in RG1, where the water table is lower. The fertilized AR2 also has an inversion of N2O concentrations during

spring that seems entirely unrelated to water table depth. In autumn, the N2O at depth declines as the water table depth increases, and the patterns in the two blocks in RG1 seem somewhat opposite of one another even though the water table is just about identical. With so much apparent variation in how N2O concentrations vary with the water table, how do these results all relate to capillary fringe as the N2O source? What am I missing? I suppose that this comment, in combination with my new response to the authors response 46, suggests an opportunity for the authors to revise and expand their presentation (and perhaps interpretation) of these results. It would be a very satisfying read if the paper spent a little more time on the subtleties in the links the graphical analysis identification of the significant drivers of surface N2O fluxes to the spatio-temporal dynamics presented in figures 3-6 (these are most of the data being presented in the paper, so feel free to give them more attention in the text).

Line 501: I'm not sure why water table dynamics are invoked in the discussion of N2O emissions at AR1, if capillary fringe was not a significant predictor of surface N2O fluxes in the graphical analysis, and, as noted previously (and in my comment on the authors' response 46), surface fluxes were already elevated before the rise in WT depth? A simpler explanation consistent with the graphical analysis might be that topsoil was the source of surface N2O fluxes in AR1 during autumn, apparently independent of water table depth. If this discussion is intended to be descriptive only of dynamics within the soil (and does not have any relation to surface N2O emissions), that needs to be made clear.

Line 528: It would be helpful to have a definition of "dead-end pores"

Line 567-559: These are all the kinds of statements that need to be made much more qualitatively in the absence of clear statistical support, and issues with experimental design and strength of statistical support for conclusions need to be highlighted in the conclusion.

Line 560: remove the comma between "hypothesis" and "that"

Line 561: my impression from the results is that NO3- reduction coupled with FeS2 reduction is likely to be a trivial mechanism of N2O production in these soils (a conclusion apparently supported by the related manuscript that has been submitted elsewhere)—this statement makes it sound as though it could be non-trivial.