Response to comments by Y. A. Teh (Referee)

We thank the reviewer for his supportive evaluation, insightful comments and questions. Addressing them will strongly improve the manuscript.

GENERAL COMMENTS

This is a creative and interesting process-based experiment that uses different aggregate treatments (i.e. micro- versus macro-aggregate dominated) and plant-soil treatments (i.e. a gradient of "plant influence," from rhizosphere to detritus-affected soil to plant-free soil) to determine how differences in soil structure and various levels of plant influence potentially influence N_2O dynamics in soil. The factorial experimental design is powerful because it enables the investigators to assess not only main effects, but also evaluate the potential importance of synergistic effects among different treatments. Overall, it is my view that this paper was clearly written, with a well-justified experimental design, and a logical analysis of the data. The introduction to the paper clearly explains the basis and wider significance of this research, while the methods section explains the overall approach taken with clarity. The results section documents the main findings of the work succinctly, while the discussion takes a reasonable (and not overly speculative) approach to data interpretation, informed by the authors' grasp of the current literature. The investigators' comprehensive measurement of a range of environmental parameters is to be commended and enables them to make logical inferences about the role of different treatments and environmental factors in regulating N_2O dynamics during different parts of the simulated water cycle. In particular, the investigators make good use of redox potential measurements to evaluate how changes in redox/O_2 availability could be driving N dynamics along the "plant influence" gradient that they have created in the laboratory.

However, while I am generally supportive of this research and believe it will make a valuable contribution to the wider body of knowledge on this topic, I do have a few general remarks that I believe need to be addressed before this paper can go forward to publication. First, I think the authors need to be open and transparent about the potential limitations of their research. For example, the soil structure treatments represent two extremes (large versus small aggregates), whereas in reality micro- and macro-aggregates would be mixed together. The authors need to explain how their experimental treatment could relate or correspond to real-world conditions, drawing if possible on pre-existing field or laboratory data (see points 1 and 5 below).

R I: We agree that including a discussion of the implicit limitations of our experimental approach with respect to natural conditions will contribute to a better evaluation of the results of our study, and we thus will include this in a revised version of the manuscript.

By investigating two pedogenetically well-defined aggregate size fractions (4000 – 250 µm and 250 – 0 µm; Tisdall and Oades, 1982) separately – but with soil structure kept similar by replacing the removed fraction by inert material of the same size - , we aimed at evaluating the individual potential of these fractions to offer conditions for the soil microbial community to form N_2O. Following the reviewer's suggestion, we propose to include a discussion of how these conditions relate to real-world conditions as follows. As detailed in our response R1 below, these two size fractions represent significant "components" both of our investigated original soil and of most other soils. However, we intentionally excluded interactions between the two soil aggregate size factions to assess the individual potential of each faction separately. Therefore we can neither assess any interactions between large and small aggregates, nor such with soil structures larger than 4mm, which all may also be important for N_2O emissions under natural conditions. Since we have no data related to this, we prefer not to speculate about such effects in our paper.

Likewise, the authors need to be clearer about the limitations underlying their rhizosphere (Salix) treatment. It is difficult to generalise more widely about the effects of plant rhizospheres on N dynamics without examining a range of different plants (including single and multi-species mixtures), in order to tease-apart individual species effects from generic rhizosphere effects (see point 6 below); I think it is important, in the revised version of this text, that the authors acknowledge this limitation and spend a bit more time exploring what they believe could be more widely generalisable from their study, rather than what is species-specific.

R II: For a reply the reader is kindly referred to R6

Second, I do not believe that the authors have fully exploited their experimental design in the analysis of their data, and sincerely believe that more could be done to examine these data in greater depth. For example, as mentioned above, one of the strengths of a factorial experimental design is that the investigators can establish if there are synergistic interactions among different experimental treatments (e.g. aggregate X rhizosphere effects). However, the investigators do not appear to have examined if interactions among treatments occurred, or at least these findings are not reported if these tests were conducted. Moreover, I would suggest that the authors try more complex multivariate models to analyse their data; for instance, using approaches such as analysis of co-variance (ANCOVA), generalized linear models, or mixed effects models. The benefit of these more comprehensive multivariate models is that they enable the investigator to establish the relative importance of different treatments and continuous environmental variables in regulating flux.

R III: We fully agree that an experiment has to be analyzed according to its experimental design. In our case, this includes the interaction of aggregate size and soil treatment (unamended, litter addition, plant presence). We in fact have included this term in all ANOVA models, but failed to report the results when the term was not statistically significant or only weakly significant. We will fix this in the revised version. The structure of our experimental treatments is not hierarchical so that no mixed model is required. Such a model would only be necessary if one would analyse the

time series data, i.e. if one had several values per microcosm. We have considered this but decided not to do so, for the following reasons:
(1) our focus was on the *average response* during distinct phases that we have identified in our time series, in particular during "hot moments" after wetting; working with average time-series data provides an answer to hypotheses about whether total emissions during this period, for example, differ between treatments; in other words, our hypotheses were about cumulated fluxes during a period, and we therefore carried out these analyses at this level.
(2) the processes we observed are extremely dynamic; fitting a full time series model would almost certainly have resulted in significant time x treatment interactions – such effects would primarily be driven by the peak values of e.g. N_2O emissions after wetting; whether treatment differences for these single measurements reflect true differences in time and extent of peak fluxes is uncertain… it in fact is very likely that the true peak occurred a short time before or after these measurements, and this may be treatment specific. Again, we were not interested in whether the maximum flux occurred a bit earlier or later in time (this may not be reproducible anyways), but whether total emissions during the hot moment changed. Working with such aggregated data solves the problem of subtle shifts in emission timing, and gives extreme values much less weight.
(3) the proper modelling of the time series is very complicated: this involved heterogeneous variances (because large values scatter more) and the modelling of serial correlations (because subsequent values are not independent). On the time-aggregated scale, these problems do not occur. We also could log-transform the data to compare the treatments, which was not possible on the raw data because (a) negative values occurred due to measurement error, and (b) we were asking questions about total fluxes (e.g. grams of N_2O emitted) and not relative effects.
In summary, we agree that more complex analyses can potentially be done. However, we have deliberately focused on (1) the aggregation level that matched the questions we were asking, and (2) the aggregation level at which statistical procedures were robust. We agree that we did not document this very well and propose to address this in the revision.

Third, I agree with the first referee that the authors need to spend a bit more time clearly highlighting what knowledge gaps this paper fills. As the first referee indicates, there are already existing studies that have examined the individual effects of all the variables discussed here. In order to make this paper more impactful, the authors need to articulate how this specific study is unique or advances our current state-of-knowledge (e.g. does the factorial design add knowledge or insight?). Specific comments are provided in the section below.

R IV: We concur with both reviewers that the specific objectives of this study were not sufficiently well stated. As mentioned in our response to Reviewer 1, this aspect will be addressed. We will clarify that, while the effects of microhabitats related to soil aggregates, the detritusphere and plant-soil interactions in the rhizosphere on N_2O emissions from soils have been studied individually, little is known about their relative

effects and interactions. In our mesocosm study, we investigated this aspect for the hot moments of N_2O emissions from floodplain soils during the drying phase after flooding. In particular, aggregate size effects have not been investigated in this context (as stated on lines 79f). A particular novel aspect of the study is the minimization of the potentially confounding factor “soil structure” by mixing a given aggregate size fraction with inert material replacing the removed smaller or larger fraction. As stated on line 71ff, previous studies employing isolated aggregate size fractions have provided partially inconsistent results possibly linked to some extent to the changes in soil structure by aggregate separation.
The better specified objectives and novel aspects will be included in the introduction of the revised manuscript.

SPECIFIC COMMENTS
1. Lines 136-137: For experimental purposes, the investigators have created quasi-artificial system conditions, with treatments either containing macro- or microaggregates. While I fully understand why this was done, it would be useful to understand (even qualitatively) how close or far from reality these treatments are. For example, what was the proportion of macro- and micro-aggregates under natural conditions?

R1: The original floodplain soil consisted of 18.5 ± 4.6 % aggregates smaller than 250 µm and 81.5 ± 4.6 % macroaggregates (mean ± sd; n = 10). We composed our model soils of a 1:1 mixture of isolated aggregates and inert matrix material. This is different from the original soil composition, but well within the range of published top soil aggregate size distributions (e.g. Cantón et al., 2009; Gajić et al., 2010; Six et al., 2000). 50% microaggregates may be more than what is found in most natural or agricultural soils. Nevertheless, we chose to use equal amounts of small and large aggregates to be able so separate effects of aggregate size from effects of aggregate amount (soil mass). To reflect these reasonings, we propose to discuss the distribution of small and large aggregates in the original soil (material and method section of the revised manuscript). The discussion of relevance would be added to the discussion in section 4.1 and in the conclusions. For additional considerations on the effect of flood disturbance on small-scale heterogeneity and dynamics of aggregate size distribution see R5 below.

2. Line 173: Clarity of expression; consider revising this section to read “The mesocosm experiment had a factorial experimental design consisting of two factors (model soil and plant-soil treatment), with the first factor containing two levels (macroaggregates, microaggregates) and the second factor containing three levels (unamended, litter added, plant present). This experimental design resulted in six treatments, each replicated six times.”

R2: The authors concur with this remark and will adjust this part accordingly

3. Line 179-180: What was the rationale for autoclaving the leaves? Under natural conditions, these leaves would contain their own microbial community which could contribute to N_2O dynamics, and autoclaving means that the results will be biased towards the activity of the soil community (or, spore-forming phyllosphere microbes able to resist the effects of autoclaving).

R3: Since we specifically wanted to test the effect of additional labile C available to the N_2O producing or consuming soil microbial community, we decided to eliminate, or at least reduce the effect of and interaction with the phyllosphere of the collected leaves by sterilization. We are aware that this introduces a certain bias. However, so far there are no direct effects of the phyllosphere community on N_2O production described in the literature. The only role of these organisms in plant-atmosphere interactions reported in the literature is in capturing/consuming methane and/or volatile organic carbon compounds (Bringel and Couée, 2015). On the other hand, we cannot say anything about potential effects of interactions between the phyllosphere and soil communities on N_2O production/consumption. These remarks will be added to the discussion section of the litter effects, 4.2., in the revised version of the manuscript.

4. Lines 221-232: Further detail on the statistical analyses are required here. For example, what were the independent variables used in the ANOVA? Did the model include interaction terms? Given that sampling was conducted over different periods of time, did the authors use a repeated measures ANOVA, to account for the effects of time?

R4: The independent variables for the two way ANOVA were SOIL TREATMENT (unamended, litter addition, plant presence) and AGGREGATE SIZE. The ANOVA model also included interactions, which were indeed significant for some of the parameter. However, we did not report the cases where the interaction was not or only weakly statistically significant. We will address this in the revision.
Our hypotheses were related to total fluxes during hot moments, which is why we did not analyze the time series but aggregated data. The rationale for this was already explained in detail above (R III).

5. Lines 300-353: This is an interesting and well-written part of the discussion. However, I do think that this part of the discussion could be improved by trying to link back the findings from the experiment to natural conditions (see point 1). For example, under natural conditions, what is the relative distribution of macro- or micro-aggregates? Based on your understanding/knowledge of the natural aggregate distributions, what patterns or processes do you think will dominate in a natural setting? While I realise this might be somewhat speculative (unless other data, such as field measurements, are available), I think it's an important talking point, as it will enable the reader to relate these findings (derived under somewhat artificial conditions) to the real world.

R5: For our assessment and evaluation of the relative distribution of macro- and micro-aggregates in our experimental soil and other soils reported in the literature see R1.
Furthermore, the frequent hydrological disturbance in floodplains creates a highly dynamic and small-scaled spatial mosaic of different aggregate size distributions. Therefore, the results on the individual potentials of differently sized aggregates to emit N_2O and their respective interactions with plant roots and litter accumulation could help to better understand the seemingly erratic spatial and temporal distribution of enhanced N_2O emissions from floodplain areas. Considering our results, one could speculate that zones with a relatively high percentage of macroaggregates would be prone to particularly high emissions during hot moments. In a revised manuscript, these considerations would be added also to the discussion in section 4.1.

6. Lines 380-406: The discussion of potential direct and indirect effects facilitated by the presence of an active root system is interesting and well-reasoned. However, I was left wondering as to how generalizable these findings are, given the wide range of traits displayed by different plants? I.e. to what extent are the trends identified here unique to Salix, and to what extent are these patterns more widely generalizable? I think it is important that the authors develop this section a bit further, in particular acknowledging this limitation more frankly.

R6: Different plant species may indeed exert different rhizosphere effects (for an overview of potential rhizosphere effects see the current manuscript lines 81 to 101). Thus, strictly speaking, this study is directly relevant only for salix sp.. However, this is an important plant genus adapted to temporary flooding and thus often found in river floodplains. While oxygen depletion by root exudation stimulated microbial respiration, discussed as one process potentially reducing N_2O emissions in our study, likely occurs in the rhizosphere of any plant, rhizosphere aeration as alternative process is restricted to plants possessing aerenchyma. However, the latter is a trait of many plants adapted to temporary flooding. It has been described also for the grass family of poaceae, or for ash, and It would not be surprising to find this trait in other Salicaceae like poplar sp. and other species of softwood floodplain forests.

References:

Bringel, F. and Couée, I.: Pivotal roles of phyllosphere microorganisms at the interface between plant functioning and atmospheric trace gas dynamics., Front. Microbiol., 6(MAY), 486, doi:10.3389/fmicb.2015.00486, 2015.

Cantón, Y., Solé-Benet, A., Asensio, C., Chamizo, S. and Puigdefábregas, J.: Aggregate stability in range sandy loam soils Relationships with runoff and erosion, CATENA, 77(3), 192–199, doi:10.1016/j.catena.2008.12.011, 2009.

Gajić, B., Đurović, N. and Dugalić, G.: Composition and stability of soil aggregates in Fluvisols under forest, meadows, and 100 years of conventional tillage, J. Plant Nutr. Soil Sci., 173(4), 502–509, doi:10.1002/jpln.200700368, 2010.

Six, J., Paustian, K., Elliott, E. T. and Combrink, C.: Soil Structure and Organic Matter, Soil Sci. Soc. Am. J., 64(2), 681, doi:10.2136/sssaj2000.642681x, 2000.