Referee comments 2 RC2

General Comments

The authors use flux observations from 101 chamber plots and a dataset of environmental drivers to estimate daytime CO2, CH4 and N2O fluxes across a study site in northern Finland over July 2018. This dataset is unique and the topic is certainly within the scope of the Biogeosciences. I think the manuscript could make a valuable contribution to the literature on high latitude greenhouse gas fluxes. However, I feel the article requires major revisions to address some key pitfalls before it can be published.

While there are a large number of sample plots, the lack of temporal replicates concerns me. I understand the difficulty and expense associated with Arctic research, so I do not feel this issue alone should disqualify the manuscript, but this limitation should be discussed more in depth. The authors briefly address this but I feel it necessitates further explanation within the text itself.

Response: Thank you for this feedback. We agree that in an ideal world, we would have study designs covering both the fine-scale spatial heterogeneity as well as the temporal variability, complemented by more controlled experiments to verify the drivers of change. In this study, however, we were primarily focusing on the spatial variation, as we had identified the lack of extensive spatial study settings from the literature. Doing the GHG flux chamber measurements alone took one month for 3 researchers, and unfortunately we did not have the resources to continue this throughout several months. We will add sentences to the main text emphasizing that spatial study designs can be used to infer correlations between variables, but correlation does not imply causation. We will also discuss if and how the relationships that we observe across spatial gradients compare with those observed in time-series studies where causal relationships can be more easily observed.

- Sampling spanned only two days, between 10 am to 5 pm. Given this - I have a hard time believing “the spatial variation in our plots covered most of the temperature variation during the growing season” without a more thorough discussion.
  - Sec S1 gives mean air & soil temperature for the chambers during observation and over the study period. I would like to see these broken down in more detail, **with soil moisture too**. Perhaps as a boxplot in the supplement?
  - I assume these samples were collected under clear weather conditions. Fluxes during and after any rainfall events would be quite different. Was there much rain in July 2018? Perhaps rainfall days should be excluded from upscaling? Is it possible an upland site that is otherwise a sink could shift to a CH4 source during/after rainfall?

Response: Thank you for pointing out the importance of the temporal representativeness of the data. The sampling spanned a month, from July 1st to August 2nd. The snow melts in May-June and plants reach their maximum biomass in July-early August, after which the autumn and senescence slowly start. We assume that clarifying this misunderstanding likely solves part of the issues raised by the referee in this comment.

We will add the following description to the Supplement: "Mean soil moisture was 27 % during the CH4 and N2O flux measurements and 24 % during the CO2 flux measurements. Mean July soil moisture between 10 am and 5 pm was 30 %. Note that not all flux plots had continuous temperature and moisture loggers; this might thus explain some of the differences between the means."

We think that the comment about rainfall events is very important. Unfortunately, we cannot say how much the GHG fluxes change after rainfall because we do not have measurements from the same plot before, during, and after rain. We will acknowledge this in the Discussion: "Rainfall events might also increase soil moisture levels and activate processes related to methanogenesis, photosynthesis and respiration as well as nitrogen cycling. While our soil moisture predictions should capture these variations in soil wetness, we only made measurements once per plot under clear conditions and do not have information about how GHG fluxes might respond to rainfall events. We might thus underestimate some of the instantaneous and longer-term changes in GHG fluxes during and after rain. "

And in the Supplement:
"Measurements were made under clear weather conditions but it also rained during the study period. Rainfall can impact the soil moisture conditions and thus GHG fluxes. It rained on 8 days during July 2018, and three of the days had heavier rain (>8 mm per day, FMI 2018). We made flux measurements during two of these days because during the measurement time, the conditions were sunny. Nevertheless, the three heavier rainfall days had clear but small impacts on soil moisture (volumetric water content (%) increased by 0.01-0.08) and it took approximately 0-3 days for soil moisture to return to the preceding soil moisture level after the rain (Fig. X). Our data show that the range and mean of CH4 flux is similar both in the plots measured during or 1-3 days after the rain and during other days, suggesting that rainfall events did not have a major influence on our results. The mean CH4 flux during or 1-3 days after the rain was -1.8 and range -4.7 and 0.2 mg C m-2 d-1 (n=14), and during other days -1.5 (mean), and -4.9 and 0.1 (range) mg C m-2 d-1 (n=72); note that wetlands were not considered in this comparison because of their uneven distribution during these time periods. In our upscaling framework, we control for the rainfall events as the GHG flux predictions are based on bi-hourly soil moisture and temperature maps that should reflect changes in soil moisture conditions after rain. "

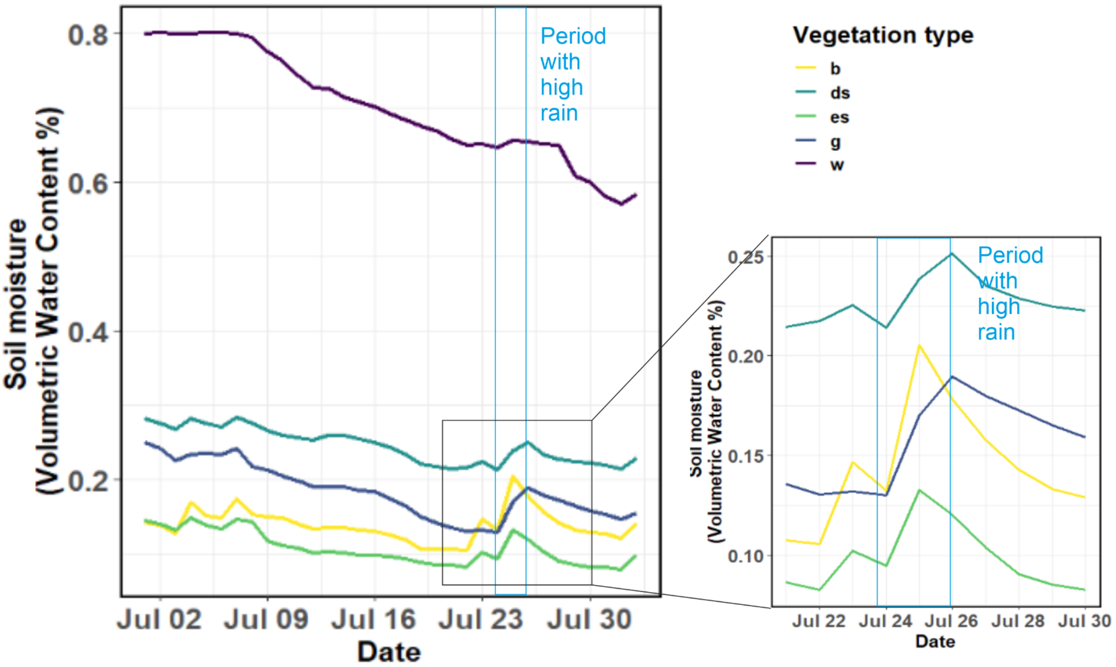


Fig X. A figure showing the soil moisture variation during the study period from 5 example plots representing the vegetation types. The subplot shows how soil moisture changes after the rain. Vegetation types are b=barren, ds=deciduous shrub, es=evergreen shrub, g=meadow, w=wetland.

Perhaps I missed it but - How many plots there were for each vegetation type?

- Were samples sizes even between types? Weighted by spatial coverage?

Response: The sample sizes can be found in Table S2. The sample sizes were not even between vegetation types, rather they represent the spatial coverage of each vegetation type. We will add a sentence about this to the main text.

I am concerned by the use of regression forest methods a dataset of this size. With 10 inputs, but only 101 flux samples (no temporal), it seems to me these models are severely over parameterized. I doubt that there are sufficient training samples for the models to adequately parse out the functional relationships in 10D feature space. It might be beneficial to consider pruning your model - you could use the feature importance to inform your choice of which variables to keep/remove. This would likely result in a more robust model that is less likely to produce spurious results.

- Random forest models are poorly suited for projection, often performing worse than simple linear regression (Hengl et al. 2018).
  - Hengl, T., Nussbaum, M., Wright, M., Heuvelink, G., & Graeler, B. (2018). Ran dom forest as a generic framework for predictive modeling of spatial and spatio temporal variables. PeerJ, 6, e5518. https://doi.org/10.7717/peerj.5518
- Looking a the partial dependence plots, the support vector machine appears to produce more reasonable results and it would be nice to see some discussion of why.

- I would like to see the authors incorporate a simpler method like ordinary least squares regression to their ensemble. I would also like to see a breakdown of the upscaled estimates for each model, in addition to the estimates for the ensemble median.

Response: In our GHG flux models, we had 101 samples and 8 predictors. While we have analyzed the model predictive performance in a rigorous and standard cross validation approach where the model parameters have also been tuned in a way that should mitigate overfitting, we understand the concern raised by the referee and agree that we are using complex machine learning models with a relatively small sample size. We will test how a generalized additive model, a simpler regression model, works in our modeling framework and compare our results with that. We will also calculate model fit, i.e., how well the models predict to the model training data to evaluate potential model overfitting. We will explore the possibility of removing predictors in our modeling framework. However, we feel that the fact that we have a similar set of predictors for all the GHG fluxes is also one of the key strengths of this study because it allows for a comparison of variable importances and response shapes.

Regarding the partial dependence plots, tree-based approaches often find thresholds in the data which are reflected as "wigglier" response shapes. This is related to the nature of the tree-based models as they split decision trees based on rules that can create these thresholds in the derived relationships. SVMs create smoother responses as the model is not based on decision trees; SVMs map the data into a high-dimensional space and build a hyperplane to separate the data and estimate smoother relationships. While we agree that very "wiggly" response shapes produced by the tree-based approaches that show no clear overall sign of positive/negative direction are highly uncertain and not suitable for large-scale extrapolations, SVMs also have their own strengths and limitations. For example, they might predict unrealistically high fluxes if the models need to extrapolate as the responses do not plateau in the same way as RFs and GBMs do. Thus, each model has its own strengths and limitations, and no model is perfect - therefore, it is generally recommended to use ensemble models in predictive efforts which we have done as well.

Taking a deeper look at the partial dependence plots, we want to highlight that most of the"wiggliest" partial dependence plots are found for variables that are less important or are produced by a model that has a low R^2 (e.g., the plot between DOC and N2O flux). To acknowledge this, we will change the y axis of all the plots to have a similar scale for each of the response variables as suggested by the referee later in the referee report. This way some of the wiggly response shapes with minor variable importance will likely only show a straight line in the partial dependence plot. We will add the following text to the Fig 5 caption so that the reader understands why the response shapes are different:
"RFs and GBMs are based on decision trees, where trees are split based on a certain threshold in the data, which can be seen as thresholds in the partial dependence plots as well. SVMs map the data into a high-dimensional space where a hyperplane is fit to separate them, creating smoother response shapes."

We will also add a figure to the Supplement showing the GHG flux predictions with all the three models.

Specific Comments

Introduction

Line 60: “and they have different spatiotemporal dynamics with each other and compared to CO2 fluxes” - reads weird, consider rewording? “All three gasses have distinct spatiotemporal dynamics.”

Response: Thank you for this and all the other language suggestions below. They are extremely helpful.

Line 66: Close parenthesis

Response: We will close the parenthesis.

Materials and Methods

Line 93: above *a* mountain birch

Response: We will change this.

Line 105 - 106: “101 GHG flux measurement plots and 50 to 5280 plots with other environ mental data” - this is confusing? 50 to 5280 plots? please explain better.

Response: We will add more details to Table S1 and change this to:

“Our study design covered an area of ca. 3 x 1.5 km and consisted of 101 plots with GHG flux measurements and their supporting environmental data. To produce continuous maps of soil temperature, moisture, vegetation type, biomass, soil C/N, soil carbon stock, and dissolved organic carbon, we utilized an extended dataset where some of the variables were measured only from 50 plots while others were measured from close to 6000 plots (Table S1). “

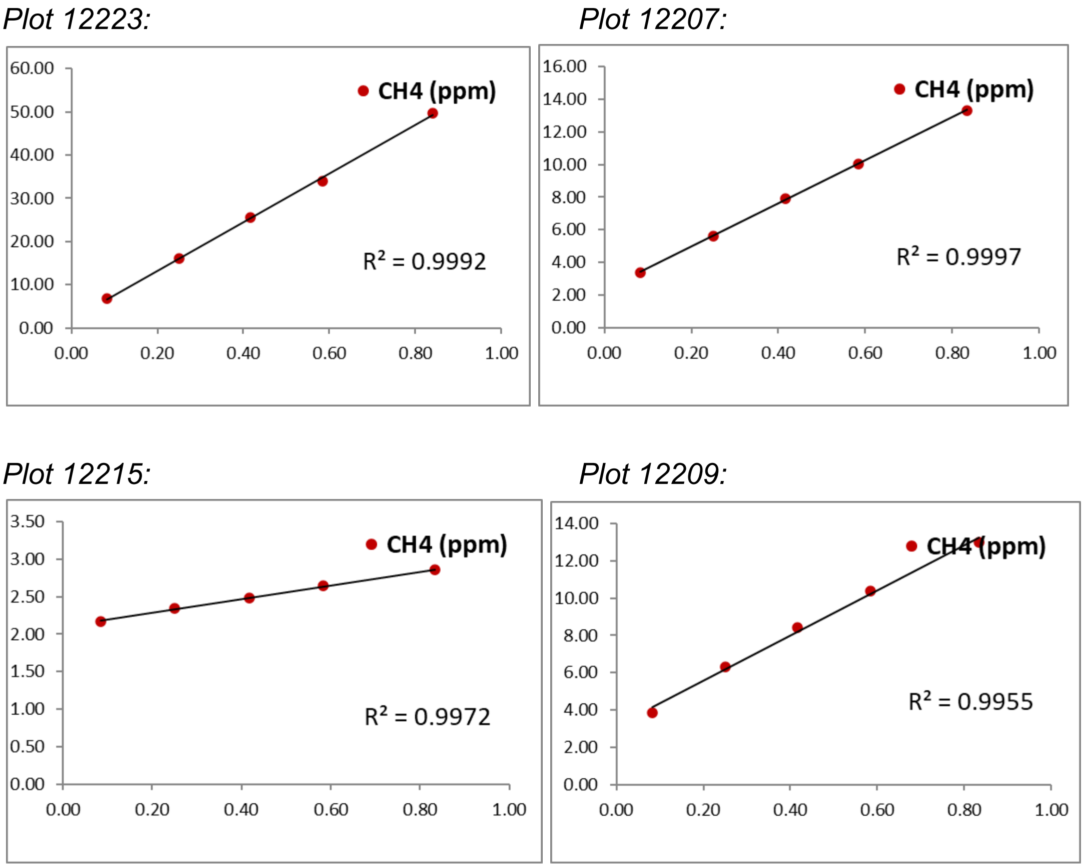
Line 121: Consider rephrasing - from Table S1 it looks like most factors had near complete coverage, so maybe say something like: “Environmental conditions explaining these GHG fluxes were measured at each plot. Most environmental variables had near complete spatial coverage; missing data were filled using the environmental predictions”

Response: We will change this.

Line 168: “Five gas samples were taken within a 50-min enclosure time” - this seems like a very long sampling interval? Are you concerned about heating within the chamber during the 50 min closure time disconnecting processes within the chamber from ambient conditions? Or about underestimating fluxes from high emitting wetland plots due to a reduction in the gas concentration gradient between the soil and the chamber head space? What was the rational for using this long sampling interval?

Response: This is a good point. For CO2, we used a 90-second measurement time, so this potential issue only applies to N2O and CH4 fluxes. We will add the following text to the Supplement:

“We hypothesized most of the N2O fluxes and CH4 uptake fluxes to be small in this landscape dominated by upland tundra, and therefore used a 50-min chamber enclosure time to detect small changes in these concentrations (for a similar closure time, see e.g. (Marushchak et al. 2021; Voigt et al. 2017) . We used an opaque chamber, covered by space tape that reflects the sun, and did not thus observe any clear signs of heating of the chamber. The chamber headspace temperature difference during the start and end of the measurement ranged from -2.3 to 0.5 degrees (25th and 75th quantiles). Despite the long chamber enclosure time, the relationship between CH4 concentrations and measurement time at sites with high CH4 emissions (wetlands) was linear, indicating no issues with the chamber closure time (see Fig. X).”



Plot 12223:

Plot 12207:

Plot 12215:

Plot 12209:

Fig X. Example graphs showing the development of CH4 concentrations at some of the wetland sites.

Marushchak, M. E., J. Kerttula, K. Diáková, A. Faguet, J. Gil, G. Grosse, C. Knoblauch, et al. 2021. “Thawing Yedoma Permafrost Is a Neglected Nitrous Oxide Source.” Nature Communications 12 (1): 7107.

Voigt, Carolina, Richard E. Lamprecht, Maija E. Marushchak, Saara E. Lind, Alexander Novakovskiy, Mika Aurela, Pertti J. Martikainen, and Christina Biasi. 2017. “Warming of Subarctic Tundra Increases Emissions of All Three Important Greenhouse Gases - Carbon Dioxide, Methane, and Nitrous Oxide.” Global Change Biology 23 (8): 3121–38.

Line 207-208: “We also utilized a larger dataset of 5820 vegetation descriptions from the study design to create the vegetation type map. Please elaborate on how this data and how the map was created. Was this product created by a previous study? If so we need the citation. If not, you need give a more detailed description of the methods and data used to create the map.

Response: We will edit the sentence:

“We utilized a larger dataset of 5820 vegetation descriptions estimated in the field and from aerial imagery from the study design to create the vegetation type map (for more details, see S4.1).”

Line 239: I don’t think one sentence necessitates its own sub-section. Perhaps expand on this a bit or merge it with section 2.2.3

Response: We will edit this.

Lines 266-268: I feel this is insufficient justification for the choice of models. I would like to see a bit more on why these methods were chosen and the pros/cons of each.

Response: We will add the following text to the Methods:

“These three approaches are non-parameteric and can handle linear and non-linear relationships. We chose RFs and GBMs because they utilize several decision trees in an ensemble model framework and thus avoid overfitting, have high accuracy, are highly adaptable, and are not significantly impacted by outliers. We chose SVMs because they are good at generalizing the relationships in the data.”

Line 302: Seems like a reasonable thing to do - I assume the idea is to minimize the effect of outliers on the model? Perhaps explicitly say that in the text?

Response: The idea is to normalize the distribution of errors, which is one of the assumptions in regression analyses. Non-normal distribution of residuals can bias, for example, the calculation of prediction intervals.

Line 320: Leave one out is concerning?

Response: Leave-one-out approach is a widely used cross-validation approach. We are not fully certain what the referee means with this comment.

Results

Lines 356-358: “The scatter plots of observed and predicted GHG fluxes suggest that the highest flux estimates are often predicted most poorly, but the mean fluxes in each vegetation type were predicted accurately.”

- This seems like an obvious point - empirical models will tend toward the mean of the training domain regardless of how well fit the distribution of the individual training points.

Response: We agree and will mention that this was expected.

Line 371: net *uptake* of CO2

Response: We will change this.

Discussion

416-417: "Aboveground plant biomass and vegetation type were important drivers for both which suggests a dominance of autotrophic (plant) respiration over heterotrophic (microbial) respiration." - this statement seems like a bit of a stretch? I would assume more above ground biomass also means more litter input for decomposition, and also would likely be correlated with below ground biomass » leading to greater microbial decomposition of root exudates? How strongly correlated were the input parameters?

Response: We agree with this and will remove the sentence.

Biomass and SOC have a correlation of -0.14 (p=0.16). The correlation is negative because the largest biomass with high Betula nana or Empetrum cover are often found in drier soils with low soil carbon stocks, whereas the wetlands have the highest soil carbon stocks but small-moderate vegetation biomass. We will add this and the other correlations to a new Supplementary table.

Lines 459-461: Out of curiosity, for what portion of the year do you expect these favorable conditions to last? I'd imagine some of the sinks, especially the valley bottom meadow would be sources during snow-melt period, and possibly again during the freeze up period in fall?

Response: This is an interesting question. Unfortunately we do not have any data covering the entire period from snow melt to thaw in this study design, but we did a CO2 flux measurement campaign in 2019 where we sampled a smaller study area three times during the snow-free season. Those data suggest that during the early growing season (mid-late June), the ecosystems were on average CO2 neutral (data published only in a Finnish-language Master's thesis; https://helda.helsinki.fi/handle/10138/331463). We agree with the referee that the meadows are likely CO2 sources straight after snow melt or right before snow arrival. This is because during the spring there is an inflow of carbon and other nutrients from meltwater streams that likely boost decomposition and during the autumn deciduous leaves of graminoids have senesced and only soil respiration is active. Interestingly though, the thesis suggested that across the smaller study design, Reco decreased more than GPP towards late summer; however, we did not capture the freeze up period in our sampling. A new year-round eddy covariance tower will be set up in this landscape in 2024 which will provide more insight on the temporal dynamics of this ecosystem.

Line 503: The gasses themselves do not act as sinks/sources, consider rephrasing.

Response: We will change this.

Line 510-511: "evergreen or deciduous shrub expansion may increase or decrease the growing season GHG sink" - consider switching to "deciduous or evergreen shrub expansion may …" to better get your point across.

Response: We will change this.

Line 536: perhaps a bit of a stretch to say "all the main vegetation types"?

Response: We will say "most of the main vegetation types" instead.

Line 543: would the temporal variability of soil moisture and temperature contribute to the difference too? What was the variability like over the period - relative to the sample days? It would be nice to se a time series of soil temperature and moisture (averaged by vegetation type) over the July 2018 study period - with the sample dates/times highlighted for reference.

Response: We controlled for the variability in soil moisture and temperature over the study period in our upscaling approach, thus this has been considered in these average flux comparisons. We will make sure that the new manuscript has time series graphs of soil moisture and temperature.

Line 556: or models having too many parameters …

Response: We will mention model structures too.

Line 575-577: An important point, well put!

Response: Thank you!

Figures

Figure 1: Plots a. and c. color schemes are misleading - using the same colors to show very different phenomena. I suggest changing the color for the vegetation maps to one better suited for discrete qualitative data. For plot b. the chambers should be color-coded by vegetation type so the reader can better see the spatial distribution of each type. Additionally, it would be helpful if the figure caption (or somewhere else in the text) said how many chambers there were for each vegetation type. For the numeric data in c. soil temperature and annual soil temperature are the colormaps are inverted relative to the other plots, which also makes things a bit unclear at first glance.

Response: We have picked one color scheme that we use throughout the text that has colors that are clearly distinguishable from each other and can easily be read by color-blind readers (or black and white paper) as well. We have inverted the color scales of temperature maps so that blue reflects cold conditions. We will test if changing the color scheme for the vegetation type makes the figure easier to read. We will add the number of chambers for each vegetation type to the figure caption. We will also color the points with the vegetation type information. We will also clarify what the vegetation type legends refer to, as this was not clear for referee 1.

Figure 3: Could you make the boxplot larger so they're easier to read and consider excluding the points that are within the boxes - makes for a confusing/overly complex boxplot. Additionally:

- You could set the y-axis limits for GPP and ER to the same values for a more direct comparison.
- You have a "-0" on your boxplot for biomass
- Maybe just keep the most important plots and move some to the appendix to save

space?

Response: We will remove the variables that were not used in the model (pH, nitrogen stock) and make the boxplots bigger so that they are easier to read. We will also set the y-axis limits for GPP and ER to the same values and change "-0" to "0".

Figure 5: Are these importance values normalized to a 0-1 scale?

- Should the bars sum to 1 for each model?
- Why is the importance for N2O so low for SVM across all features
- Please use a common y-axis range across the subplots for easier comparison
- Any estimate of uncertainty in these feature importance estimates? e.g., for a RF model you can get the importance of each sub-model and use it to calculate a 95% CI over around the RF feature importance values.

Response: The bars have been calculated using the permutation approach with the idea that if we randomly permute the values of an important predictor in the training data, the training performance would degrade (since permuting the values of a predictor destroys relationships between that predictor and the response variable). We did not normalize them between 0 and 1 and the bars should not sum to 1 but each bar is theoretically limited between 0 and 1 because the approach compares R^2 values.

We will add the following text to the main text:

"The importance for variables explaining the N2O flux is low because the model predictive performance is close to random. Variable importance scores were calculated using a permutation approach with the idea that if we randomly permute the values of an important predictor in the training data, the training performance would degrade. However, for N2O fluxes this random permutation had minimal effect on the predictors."

We used 100 simulations to calculate 100 importance scores which were eventually averaged, but we will show the variability in those importance scores too. The differences in importance scores across the models also provides information about uncertainty.

Figure 6: It seems to me this plot highlights how regression trees (RF and GBM) models are poorly suited for this type of analysis, particularly given the small sample size. The noisy response curves they generate are because regression trees are **treating each terminal node in a tree as a discrete data point to match rather than a continuous response function to fit.** The idea behind a random forest, is that averaging many of these over-fit trees will emulate the desired response function (*only within the bounds of the training data*). However, the there do not appear to be sufficient samples for the RF model to be able to do that.

- Are these partial dependence yhat values in the same unit/scale for each predictor? If so it would be useful to have each y-axis on the same scale to see the relative magnitude of the partial dependence by variable.

Response: We are not sure if one can conclude from these graphs that the response curves for RF and GBM are noisy. After all, most of the highly important variables in models that performed well have relatively clear response shapes with some minor wiggliness. The thresholds that exist in the partial dependence plots are simply related to the nature of tree-based approaches as well as thresholds that make ecological sense. For example, the sharp jump between the soil moisture and CH4 flux data can be seen by a bivariate plot as well, because with soil moisture at around 60 % soils become saturated which boosts methanogenesis (see figure below). We will change the y axis scale of the partial dependence plots, this is a great point. For a longer response to this, see our response earlier to the major points raised by this referee.

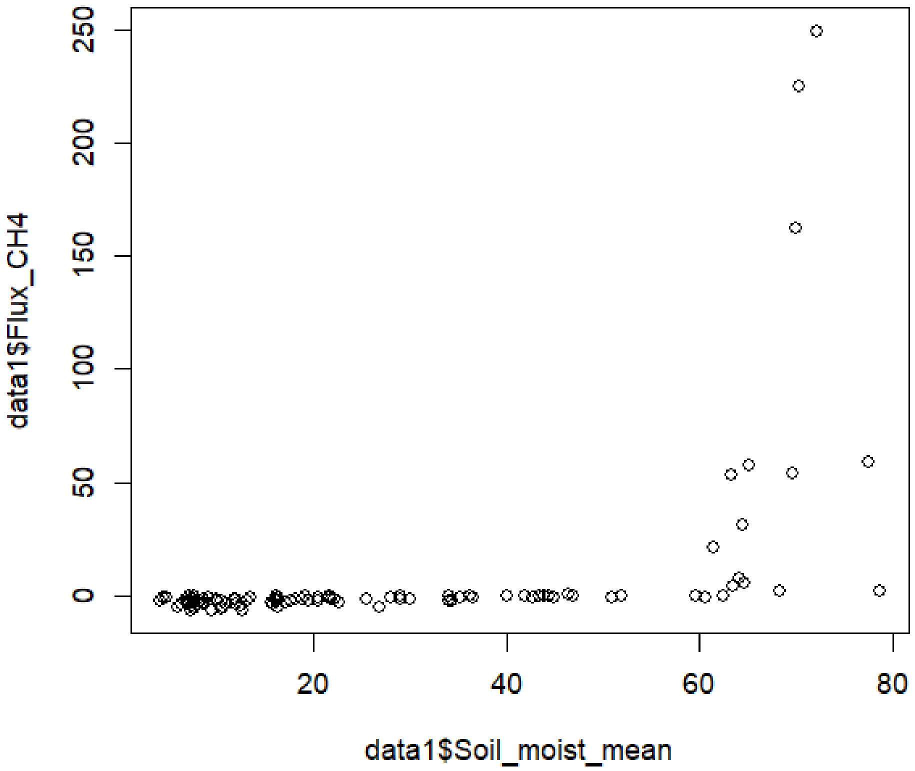


Figure. The relationship between observed instantaneous soil moisture and CH4 flux. Units: volumetric water content (%) and mg CH4 m^{-2} d^{-1}.

Figure 7: Using “strong” and “moderate” to describe the CH4 sink strength of upland areas seems odd. By area, yes they may be large sinks overall, but on a per unit area basis they are not given their relatively small magnitude compared to wetland CH4 emissions shown in Fig 8. Perhaps try rephrasing the labels in the images? Alternatively, show your landscape map here to emphasize that you’re talking on a “per landscape fraction” basis.

Response: We will change the language here.

Figure 8: I would like to see a different color for the error bars to help them stand out from the plots more.

Response: We will use a different color (e.g., grey) so that they can be better separated from the blue bars.

Supplement:

Figure S4: Needs a legend for the plots.

Response: We will add the legend.