Review of "High-resolution spatial patterns and drivers of terrestrial ecosystem carbon dioxide, methane, and nitrous oxide fluxes in the tundra"

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.

- 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?

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

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

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). Random forest as a generic framework for predictive modeling of spatial and spatiotemporal 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.

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."

Line 66: Close parenthesis

Materials and Methods

Line 93: above *a* mountain birch

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

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"

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?

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.

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

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.

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?

Line 320: Leave one out is concerning?

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.

Line 371: net *uptake* of CO2

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?

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?

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

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.

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

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.

Line 556: or models having too many parameters …

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

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.

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?

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.

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.

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.

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

Supplement:

Figure S4: Needs a legend for the plots.