Response to comments by Anonymous Referee #2
The author comments and answer are written without highlighting, while the comments of the Anonymous Referee #2 are highlighted in *cursive*.

Overall:
Scientific significance: excellent, scientific quality: excellent, Presentation quality: good
Overall this is an excellent paper.

Answer: We thank Referee #2 for the productive feedback on our study that has greatly improved the study. Please see our specific answer to the comments below. Note that we have added numbers to the comments. The line numbers below refer to the revised version of the manuscript with the track changes.

Many of my comments are requesting clarification or more details. Aside from these minor points, I have two major issues to point out:

1. *The first about the lack of a spatially explicit flux modeling for CO2 as was done for CH4 and N2O. Much of the paper is justifiable building expectations for the impacts of spatial heterogeneity after clear cutting, and it is surprisingly absent in the results and discussion for CO2. In comparison to CH4 and N2O, I would expect CO2 to be easier to model given its strong relationships to variables already reported in the gap-filling discussion. The authors could take the GPP and respiration models used with gap-filling and apply the same spatial disaggregation technique as they did with CH4 and N2O.*

Answer: We considered analysing CO_2 flux observations similarly as done for CH_4 and N_2O, but opted not to. Please see our response to Referee #1 comment 7 for the reasons for this decision.

2. *The second issue is about the methane flux results. The flux estimates from the plant-covered ditch surface-type are extremely large, almost unbelievably so. These results need to be justified and put in context of other methane emissions. Given that the areal contributions of this surface type and therefore their weights within the footprint, are so small, it could be very difficult to have confidence in these results. In addition to comparisons to chamber fluxes or other studies, I would suggest investigating the robustness of the methane surface-type model with a simulation. Generate a flux for each surface-type based on your equations 3 and 4, calculate the theoretical EC observation after multiplying by the pixel footprint weight and summing, then add some reasonable random noise. Then apply your dissaggregation model and see if you can recover the original parameters you used to generate the fluxes. This is a straightforward way to test whether your dataset is under-determined or not. If you do not have enough variability in footprints weights from surface-types to recover your simulated fluxes, then you will have to reduce the complexity of surface-types or use a longer time series of data.*

Answer: We thank the referee for this suggestion. We ran the suggested test such that one author calculated "an artificial flux data set" as suggested and then another author performed the model parameter estimation without knowledge of what were the correct parameter

values. The correct parameters along with MAP estimates from the parameter estimation are shown in Table 1 below. Since the author performing the model estimation did not know how many surface types the correct answer set had, he ran the parameter estimation with 3,4,5,6 and 9 surface types. The best performing set was ST6 closely followed by ST9. From Table 1 it can be seen that since dead wood and harvest residue and field layer and living trees have the same correct γ and δ the correct number of surface types was six in the artificial flux data set.

Table 1: Correct parameter values (columns 2-5) and the maximum a posterior (MAP) estimates of the same parameters (columns 6-9) for the artificial data set and parameter estimation performed with it. In columns 6-9 two values are given for each parameter: the first is from the parameter estimation using six surface types and the latter from using nine surface types.

Surface type	α	β	γ	δ	α (MAP) ST6 / ST9	β(MAP) ST6 / ST9	γ (MAP) ST6 / ST9	δ (MAP) ST6 / ST9
-	2.0	0.0			2.1 / 2.1	$1.4 \cdot 10^{-5}$ / $6.9 \cdot 10^{-5}$		
Dead wood			-0.1	0.0			-0.12 / -0.12	$3.9 \cdot 10^{-4}$/ 0.0017
Harvest residue			-0.1	0.0			-0.12 / 0.30	$3.9 \cdot 10^{-4}$/ 0.0011
Exposed peat			0.1	0.3			0.10 / 0.13	0.40 / 0.40
Litter			0.3	0.001			0.34 / 0.20	0.003 / 0.004
Bottom layer (mosses)			0.0	0.0			NA / -0.9	NA / 0.24
Field layer			-0.2	0.0			-0.25 / -0.27	$2.4 \cdot 10^{-5}$/ $1.4 \cdot 10^{-5}$
Living tree			-0.2	0.0			-0.25 / -0.35	$2.4 \cdot 10^{-5}$/ 0.0035
Plant covered ditch			5	0.0			6.31 / 6.2	0.014 / 0.0038
Ditch (water surface)			-2	1.8			-2.61 / -2.7	2.26 / 2.28

Additionally, below is a figure showing the distribution of the estimated parameters

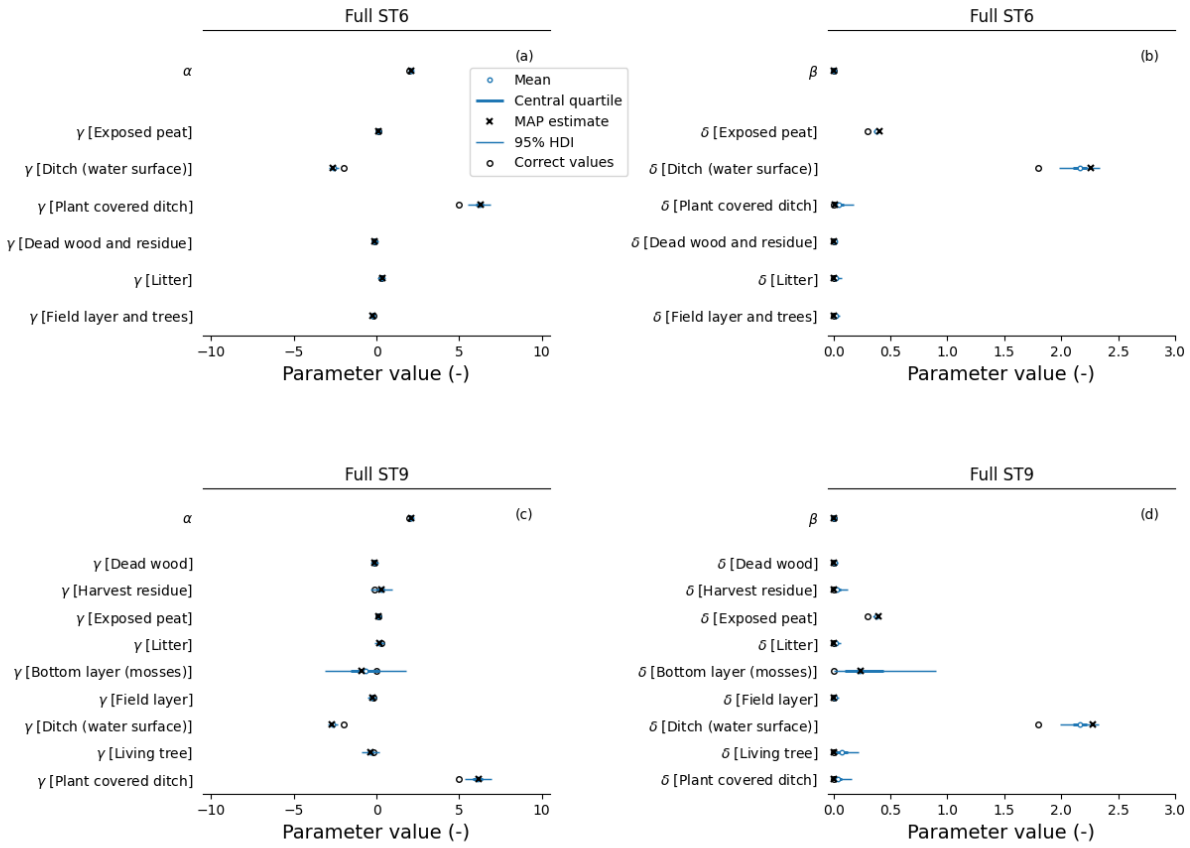

Figure 2: Inferred parameters from the artificial data set, their distribution and correct parameter values. Subfigures a-b show the estimated and correct parameters for the model with six surface types and c-d for the model with nine surface types.

In our opinion, the test showed that there is enough variability in the footprints that the modeling can be performed with the experimental data set at hand. However, it is also clear that the estimates for the both ditch types can be biased (e.g., 23%-26% for the plant covered ditch in this test).

The methane fluxes for individual surface types decreased in the revised version of the model since some of the flux is now attributed to the term with soil moisture. We are stating already that the method to calculate surface type specific fluxes is an extrapolation of the model (line 510). We feel, however, that the publication of these extrapolations is needed for later comparison against e.g., chamber measurements.

General comments:

3. *In figure 1 and throughout when color-coded landcover types are displayed: It is difficult to distinguish similar colors. The greens in particular all look the same. A more divergent color scheme would improve readability throughout the paper.*

Answer: We have adjusted the colors in the revised version of the manuscript.

4. *Model predicative performance for the gap-filling ML models and the spatially explicit footprint flux models is evaluated and reported using R-squared. Whenever R-squared is reported, the slope and intercept of the regression should also be reported. R-squared describes the variance around the fit, but the slope and intercept describe model bias which is equally important. I also suggest providing the RMSE as a more useful metric than R2 because it is in comparable units.*

Answer: We have added slope and intercept information to the flux model and gap filling ML-model comparison. The best model selection is done solely based on the ELPD-LOO in the revised version.

5. *Section 2.8:*
The methods described for surface-type modeling are the same as those used by Ludwig et al. 2024 from your introduction, and it should be cited here as well.

Answer: We now cite Ludwig et al., (2024) in section 2.8.

6. *Can you please provide some justification for your choice of prior distributions.* ▪
Please describe your tests for convergence and their outcomes. ▪
Please clarify that only non-gap-filled data were used in the surface-type modeling analysis

Answer: Our decisions on choosing prior distributions was to keep the priors as uninformative as possibly while still incorporating the little knowledge that we have of the system. The addition of θ (soil moisture) to the model makes the interpretation of the parameters slightly more challenging i.e., we cannot say anymore that the variable α is the base gas emission rate at $\mathrm{T_{air}} = 10^{\circ}\mathrm{C}$. For this reason we went with the normally distributed priors around zero mean for both α and γ and ζ. We also briefly considered using uniform priors but neglected this option as it would've meant that we believe that high values of these parameters are as likely as those near zero. For the temperature response parameters (β and δ) we went with exponential distributions as we assume that above $\mathrm{T_{air}} = 10^{\circ}\mathrm{C}$ the effect of temperature to emissions is positive. Lastly, we ensured that the values we chose for the standard deviation and rate parameters of the priors were such that the full width at half maximum (FWHM) of the prior predictive distributions is at least two times of FWHM of the observations.

The convergence checks are run by default in the PyMC sampler. Most important for us is the Gelman-Rubin statistic (r-hat). The sampler warns if r-hat is higher than 1.01 for any parameter (the source code for the convergence checks can be found in the PyMC repository https://github.com/pymc-devs/pymc/blob/main/pymc/stats/convergence.py).

Only non-gap-filled data were used in the surface type modelling analysis.

7. *Why use LOO cross validation for the surface type modeling, when you already*

have withheld data in artificial gaps created for the gap-filling ML models?

Answer: We use the whole available gas flux data sets to fit the surface type models. This means that the artificial gaps that were created in developing the gap-filling ML models are not present when we develop the surface type models. We have added clarification to lines 387-389 in the revised version of the manuscript.

"The full, non gap-filled, EC flux data sets were used in the parameter estimation i.e., the artificial gaps introduced to the flux data sets for developing the gap-filling model were not present in this parameter estimation."

8. *Figure 4 and 5: include slope and intercept on the fit depicted in panel c.*

Answer: This information has been added to the revised version of the manuscript.

9. *Figure 6: The bold line for the central quartile is hard to distinguish, can you make it bigger?*

Answer: We have increased the line width for the central quartiles. We have also made several other changes to Fig. 6 to improve readability as suggested by Referee #1.

10. Table 3: I understand that the gap-filled budgets in the second and third column are agnostic to the area and make-up of the footprint. How are the surface type modeled fluxes summarized to comparable numbers to the gap-filled EC data, given that each observation has a different distribution and weight of surface types? The modeled fluxes can be weighted by footprints before summarizing to a budget, but due to gaps, there are timepoints without footprints. It would make more sense to me to use your surface-type models to calculate the budgets for the entire domain in your Figure 1, and then similarly apply the gap-filled time series of fluxes to the same area when summarizing, rather than reporting on a per area (ha-1) basis. By controlling the areal extent of this comparison it might also reveal interesting agreements or discrepancies between the surface-type model budgets and the footprint-agnostic gap-filled budgets.

Answer: This information was missing from the previous version of the manuscript. In the revised caption for Table 3 we are stating that the modelling approach estimate is calculated with the share of each surface type from the whole clearcut area not from individual footprints. If we understood correctly what the referee is asking, the revised Table 3 has the data that is suggested here. The per area fluxes can be converted to the whole clearcut area flux by multiplying with the clearcut area (ca. 6.1 ha).

11. *Section 4.1 first paragraph:*

The spatial heterogeneity is generally put in context of similar ecosystems and other clear-cutting studies. But what is lacking is a quantitative comparison of the magnitude of these fluxes determined here (figure 7) to other studies. For example, is your exposed peat flux typical of peat ch4 fluxes? While I am not

surprised by a slight uptake of methane in some surface types, it is surprising to see methane uptake in the ditch surface water. Similar features in polygonal tundra are large methane sources. The methane flux from plant covered ditches, the vast majority of all methane at this site, is alarmingly large, as in, it is similar to methane fluxes measured by eddy covariance at active landfills in warm climates. This result needs to be put in context of other fluxes and justified.

Answer: The fluxes from different surfaces shown in Figure 7 are different from measured results. Open water ditch should be large CH_4 source but it's not reflected in our results. One reason is the main ditch which contribute most of CH_4 emissions in our site was identified as plant covered because of vascular plants growing near the ditch. This also explained the large emissions from plant covered ditch we got. The CH_4 fluxes from exposed peat varied in our study site based on our chamber measurements, depending on the water table in the location. It's difficult to quantitatively compare Figure 7 with measured results, because they are calculated by setting specific surface-type contribution to 1 which is a considerable extrapolation of the model. Please see also our answer to further comment 6 of Referee #1 for why we can't report a percentage contribution of different surface types to the overall flux.

The key information we bring is to identify the relative important surfaces which have high emission potentials, which help to know which surfaces should be considered for conducting measurements.

12. In table 2, you set up an investigation of scenarios to determine the level of complexity to use in the spatial disaggregation of fluxes. This is a great tool for supporting the robustness of your surface-type model results. You present results from the best model of the set described in the table. I would like to see more results on all scenarios. Specifically, how do the surface type flux estimates change in each version in table 2? In two of the five versions, your highest flux type is lumped with your lowest flux type, and discussing how the fluxes turn out in these scenarios would help provide confidence in the model results.

Answer: In the revised version of the manuscript we are reporting the estimated parameter values for full model with θ for 3,4,5 and 6 surface types in the supplement. In our opinion the results seem to provide more confidence that the ST estimates for the best model are coherent given the limited amount of data we have. We have added the following paragraphs to the results section on lines 535-545.

"Fig. S9-S12 show the estimated parameters for the full θ models for the other number of STs. Interestingly, for CH_4 when the two types of ditches are lumped into one ST, their γ estimate is close to zero (Fig. S9 and S11) whereas when the ditches are considered as separate STs the estimated γ for the plant covered ditch is the highest and the γ for the ditches with water surface is the lowest which is the same behaviour what we see in Fig. 6 for the best model.

The parameter estimates between different number of STs for N_2O models differ more than for CH_4 models. For example for ST6 (Fig. S12) the highest γ MAP estimate is for dead wood and residue whereas the γ for the field layer and trees is the smallest. The γ estimates for ST5 (Fig.

S11) seem to also emphasize the role of litter and dead wood and residue as high N_2O emitting surface types. It should be noted that for all other number of STs the living trees are always lumped together with some other surface type or types. It might be for this reason that the full θ no δ ST9 model outperforms the full θ ST6 model for N_2O models but not for CH_4 models (Table S1)."

Specific comments:
Line 88: Need space after period at the end of sentence
Line 107: Missing word. "[The] likely reason for this…
Line 247: missing space in citation for (Kljun et al 2015)
Line 565: Should cite Ludwig et al. 2024 here as well.
Line 581: Typo 'emissionsdd,

Answer: These specific comments are included in the revised version of the manuscript as suggested.