We thank the reviewer for the thorough and constructive comments. This detailed feedback has helped us identify several aspects that require clarification and improvement, particularly regarding data representation, model validation, and the structure of the *Methods* section. We will carefully address all points raised and provide corresponding clarifications, additional analyses, and improved figures and tables in the revised manuscript. Our detailed responses to each comment are provided below.

RC: *It is unclear whether the sample is representative of the entire study area. The data appear to be concentrated in a few locations, which is understandable given the logistical challenges of conducting fieldwork across large wetland areas. Nevertheless, this aspect should be described in more detail in the 'Methods' and 'Results' sections; a single sentence in the 'Limitations' section is insufficient.*

A more detailed discussion of site representativeness would be beneficial, including the number of sites per land cover type, and the number of measurements per site. In Figure 4, n appears to refer to the total number of measurements, but it would also be useful to indicate the number of sites per land cover category there, as well as in Tables B1 and B2, or somewhere else. This would facilitate discussion of this limitation, e.g., the text mentions that the 'wetland, permanent' class includes only one site which.

Additionally, are sites weighted differently in the model? For example, automatic chambers likely produce more measurements than manual ones — do these sites then have a greater influence on model training? How do you account for potential site-level overfitting? Did you consider using a leave-one-site-out cross-validation approach to assess the robustness of the model in predicting new areas where no data was used for training?

Finally, could the differences between the two models (particularly at 10 m in Figure 6) over specific areas/LC types be explained by a lack of training data in these areas/LC types?

AC: We agree that the current description of the sampling coverage and data balance across sites and landcover types requires more detail and will substantially expand this part in the revised manuscript.

In the revised Methods section, we will include a more explicit description of the number of sites and measurements per landcover type, also indicating which landcover classes are represented by automatic and manual chamber systems.

Automatic chambers were deployed primarily in drier or shrub-dominated surfaces, whereas manual chambers represent wetter sedge and mixed sites. Together, these sites cover the main landcover types present within the study area, even though they are spatially concentrated in several field clusters. We will also indicate the number of sites per land-cover class in Figure 4 and summarize this information in a new supplementary table for improved transparency. This will clarify how site distribution reflects the heterogeneity of the study area and where data density is lower.

All chamber measurements were treated equally in the ML workflow, without explicit site-level weighting. Consequently, sites with automatic chambers contribute a larger number of observations, reflecting their higher temporal resolution. To mitigate potential overfitting to these sites, we used grouped cross-validation with “Site” as the grouping variable, ensuring that all data from a given site were included either in the training or in the testing subset, but never in both. This design is conceptually similar to a leave-one-site-out validation while maintaining multiple folds to preserve representativeness across sites. We will clarify this explicitly in the Methods section and emphasize in the Results that this approach provides an effective test of model robustness in predicting sites not used for training. Additionally, we will perform a sensitivity test with a full leave-one-site-out cross-validation to further evaluate generalization performance.

RC: *Mismatch between data input and resolution effect :*

The comparison between the 1 m and 10 m datasets is particularly interesting, as it reveals the differences in the two approaches with commonly used input data at these resolutions. However, this comparison potentially combines two effects: one related to the resolution itself (average over a larger area), and another related to potential differences in the data sources themselves (e.g., different acquisition date/time, different sensors...). Have you attempted to separate these two influences? One way to do this would be to aggregate the 1 m product to 10 m and apply the same workflow (e.g. for land cover, use the dominant vegetation type within each 10 m grid cell and take the mean for the other variables). This could help to isolate the effect of the resolution from that of the different data sources.

Otherwise, it would be useful to discuss this somewhere, and include a comparison of the datasets used as is done for the comparison with CALU (Figure 5), but for the two datasets at different resolutions (as is partly done for land cover in Figure 2, where important differences can be seen).

AC: We agree that the comparison between the 1 m and 10 m datasets may combine two effects: (1) the change in spatial resolution and (2) the use of different data sources. To evaluate this, we performed an additional analysis where all 1 m input layers were aggregated to 10 m resolution, and the same modeling workflow was applied using identical parameter settings. We will include this additional analysis and its figure in the Appendix of the revised manuscript and refer to it in the Results and Discussion sections.

At the same time, we will explicitly clarify in the text that the purpose of this study is to evaluate CH_4 flux upscaling using freely available datasets (Sentinel based) vs UAV/drone products at 1 m. Therefore, we will retain the main comparison between these two operationally distinct input datasets, while providing the aggregated 1 m to 10 m test as complementary evidence that supports the interpretation of model differences.

We will also expand the discussion of input-data differences between resolutions, adding a short comparison of key variables in the supplementary material, similar to the approach used in Figure 5 for the CALU comparison.

RC: *The mix of spatial and interannual analysis is somewhat confusing.*

It is unclear how the spatial and interannual components are distinguished in the study. It is not always obvious whether the analysis is spatial, temporal, or a combination of both. Although the study appears to be mainly spatial, with a single-month focus on July, it also uses temporally varying predictors only (AT, PAR and TDD over six years). Clarifying this in the text and figures (methods and results) would improve readability. The time-varying inputs are difficult to understand from the main text: which variables are dynamic and at what resolution? (See the comment about the data section below.)

Spatial and temporal accuracy should be discussed separately in the 'Results' section, or more explanations should be provided. For example, spatial correlations (mean flux per site) and temporal correlations (time series at individual sites) could be reported separately in Tables 2 and 3 to disentangle these effects and avoid sites with potential larger amounts of data dominating the analysis compared to sites with smaller amounts of data. A panel like 7B could be used to directly compare model predictions with measurements at the sites, providing a clearer assessment of spatial and temporal performance.

AC: Indeed, our study is primarily spatial, focusing on small-scale variability in methane fluxes across different wetland elements within the fixed July period for each year (2019-2024). The inclusion of temporally varying predictors (AT, PAR, and TDD) serves to capture the short-term meteorological variability among measurement dates within this single-month window rather than to represent long-term seasonal or interannual trends.

To clarify this, we will explicitly state in the Methods and Results that:

- The spatial component refers to differences among sites and landcover types within each year.
- The temporal component reflects variability among measurement days within the study period (late June-July).
- The interannual aspect is limited to comparing the same seasonal window across three years.

RC: *The data section of the Methods section needs to be restructured and expanded.*

- Section 2.2.3 (and the Materials and Methods section more generally) should be reorganised, as it is currently difficult for the reader to determine which datasets are used at 1 m, which at 10 m, and which at both resolutions. For instance, the text initially focuses on 1 m data, but then abruptly shifts to Sentinel-2 (presumably 10 m) before describing the 10 m products. References to 30 m window data are confusing and require explanation. The temporal dimension of each variable is unclear too. While some of this information appears in Table A1, Figure 3 and lines 240–247, the description remains fragmented. Providing a summary table that explicitly lists the ten variables used for each resolution, their data source, spatial resolution (1 m, 10 m or constant) and whether they are static or dynamic would certainly help the reader.

- Data processing procedures should also be described in more detail in Section 2.2.3 or in a dedicated section. For instance, how were Sentinel-2 data cleaned or filtered? Were cloud-free conditions explicitly selected for the time-varying Sentinel-2 indices? This is implied by lines 278–281, but stating this explicitly in the 'Remotely Sensed Data' section would improve transparency. Overall, providing a clearer and more detailed description of the data pre-processing and management would strengthen the reproducibility of the study.

- for the chamber data, management should also been specified. How is chamber data managed spatially? How are fluxes aggregated at 1 m or 10 m resolution — do you take the mean of all chambers within each 1×1 m or 10×10 m pixel? You mention PAR and other variables measured at chamber sites. Are these used here ? Providing this information is essential for understanding how point-scale observations are scaled to the model resolutions. Additionally, since chamber flux measurements are known to be highly variable, it would be useful to specify in the methods section whether each flux observation corresponds to a single or repeated measurement.

AC: We agree that Section 2.2.3 requires clearer organization. We will restructure this section to explicitly separate datasets used at 1 m, 10 m, or both resolutions and add explicit references to Table A1 (Appendix A), which already summarizes the predictors, their data sources, spatial resolution, and whether they are static or dynamic. We will also clarify that 30 m window variables (e.g., TPI 30 m) describe topographic context and are applied consistently at both resolutions. The description of Sentinel-2 preprocessing will be expanded to explicitly state that only cloud-free summer scenes were used for NDVI and NDWI derivation.

For chamber flux data, we will clarify that fluxes were aggregated by averaging all chamber measurements within each 1 × 1 m or 10 × 10 m pixel, and that each flux observation represents the mean of repeated chamber measurements taken during the same campaign.

RC: *I do not understand Figure 7A. According to the caption, it should show monthly estimates averaged over the entire area, but the large number of points is confusing.*

AC: Figure 7A indeed shows monthly CH_4 flux estimates for individual pixels across the study area, not a single aggregated mean. The large number of points reflects spatial variability within the domain. We will clarify this in the caption and text and indicate that each point corresponds to a pixel-level monthly mean to improve readability.

RC: *Have you considered using a "leave-one-site-out" or "leave-one-year-out" cross-validation (e.g. training on the first three years and predicting the last year) ? This could enable assessing how well these models can predict pixels/sites or time for which no data was used in the training process, as well as the uncertainties related to each model training, which are not really discussed here.*

AC: Thank you for the suggestion. Our current cross-validation already uses grouped folds by site, which partially addresses this issue. However, we will explicitly state this in the Methods and discuss how a full leave-one-site-out or leave-one-year-out scheme could be implemented in future work to further test model transferability.

RC: *Linking these fine-scale results to broader CH_4 budgets, which are usually estimated at coarser resolutions, raises questions about scalability. Why were only 1 m and 10 m resolutions considered? Would other coarser scales (50 m, 100 m, 1 km) be relevant ? The comparisons mentioned in lines 436–440 refer to models run at 0.25–0.5°. Are these results directly comparable? How could your findings be used in larger-scale budgets?*

AC: We appreciate this important point. Our focus on 1 m and 10 m resolutions was motivated by the need to bridge the gap between field-scale chamber measurements and satellite-based observations, particularly those derived from Sentinel-2 and drone imagery. These two resolutions thus represent the most relevant scales for practical upscaling of chamber data. We agree that exploring coarser aggregations (50–100 m or 1 km) would provide valuable insight into the scalability of our approach. We will note this as an outlook for future work and clarify that our results are not directly comparable to regional-scale models (0.25-0.5°), but rather provide fine-scale inputs that can support parameterization and validation of such coarse-resolution CH_4 budget models.

We will also take the remaining reviewer comments into account in the revised version, basically accepting all suggested edits to further improve clarity, data description, and consistency across sections.