Overall, I'm happy with both the authors' comments and their efforts.

However, there is one important point that I don't fully understand, but perhaps I'm missing something here.

This relates to my questions about site overfitting. The authors responded to my concerns that they were using five-fold cross-validation grouping data per site (is it really the case in the general model optimisation setup ?). Five-fold cross-validation is then supposed to prevent site overfitting, as the data of one site is either used for training or evaluation, but not both. But they also show some tests they made for CV-Site in the supplementary material. I am then wondering why the Site-CV results (Text S4 and Figure 7) are so different from the five-fold CV results, as both methods should avoid site overfitting?

I understood that 'Grouped by Year' results were not really meaningful here. However, the results shown here for 'Grouped by Site' would indicate very poor model reliability. This would change the results of the study and call into question the ability of the models to reproduce fluxes and to be upscalled over the area.

The two RC/RA items related to this are listed below:

RC: 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?
AC: All chamber measurements were treated equally in model training, without explicit site-level weighting. Automatic chambers indeed produced more frequent measurements and therefore contributed a larger number of samples, reflecting their higher temporal resolution. Automatic and manual chambers were deployed in different land-cover types, except for one mixed class (class dwarf shrubs), where both were present. Because overlap was limited, the larger number of automatic-chamber observations did not substantially bias model training toward specific surface types. To prevent overfitting to individual sites, we applied grouped five-fold cross-validation using Site as the grouping variable, ensuring that all data from a given site were contained entirely in either the training or testing subset, but never in both. This setup prevents data leakage and provides a robust assessment of model transferability to unseen sites. We also performed leave-one-site-out (LOSO) cross-validation, which produced results similar to the grouped CV; therefore, it was not used as a separate evaluation.
+
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 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: We performed grouped cross-validation using both Site and Year as grouping variables to evaluate model transferability beyond the calibration domain. In the grouped-by-site CV, all data from a given site were held out entirely in the test set, while in the grouped-by-year CV, all measurements from a given year were excluded during training. These setups are equivalent to leave-one-site-out and leave-one-year-out approaches but implemented with balanced folds to ensure stable statistics. As expected, R2 decreased and RMSE increased compared to standard v-fold CV, because environmental conditions varied strongly between sites and years. Many sites were not revisited in all years, and specific landscape classes (ex., drier) were sampled predominantly in earlier campaigns (2019-2021), while wetter surfaces were

measured mainly in later years (2022-2024). Thus, holding out entire years or sites effectively removed key combinations of soil moisture, vegetation, and temperature from the training data, making it difficult for models to predict fluxes under unseen environmental contexts. Nevertheless, the models retained the correct direction of CH4 fluxes (high fluxes at wet sites, near-zero or negative at dry ones), indicating a consistent mechanistic response despite reduced fit. These results underline a critical need for repeated measurements under comparable local conditions across years to better constrain interannual model generalization. For clarity, we summarize these diagnostics in the Supporting Information (Text S4, Fig. S7) and briefly discuss them in the revised Results sections (line 376).

And the Figure S7 :

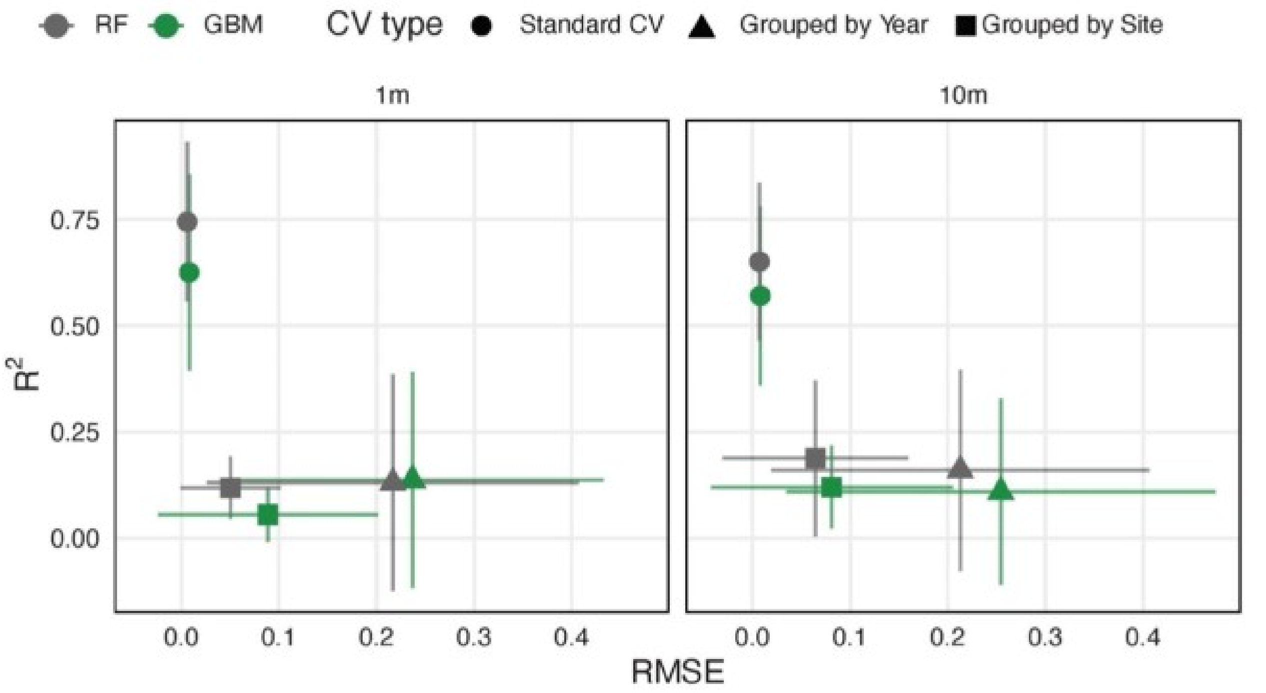

Figure S7. Model performance (R^2 vs. RMSE) for Random Forest (gray) and Gradient Boosting Machine (green) models under different cross-validation schemes (Standard CV = squares, Grouped by Year = triangles, Grouped by Site = circles) at 1 m and 10 m resolutions. Points show mean performance across folds; whiskers show standard deviation. Lower RMSE and higher R^2 indicate better performance.

citation of Text S4 : "Under a standard five-fold CV, models achieve high accuracy (R2 = 0.7-0.75, RMSE ≈ 0.06-0.07 for both 1 m and 10 m resolutions). In contrast, grouped-by-year and grouped-by-site CV produce much lower R2 values (typically 0.1-0.2) and larger RMSE (0.15-0.4)."

Minor comment :

I'm not sure Figure's titles have been changed ?
perhaps a bug ?

RC: 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.*

AC: The total number of measurements per land-cover class is already shown at the top of Figure 4. To improve transparency, we have now added the number of sites per class directly in the figure caption and included both site and measurement counts in Table B1.

--> I don't think the changes worked in the figure caption.

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 CH4 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 clarified this in the caption and indicate that each point corresponds to a pixel-level monthly mean to improve readability.

--> I think then "averaged over the entire area of interest" should be deleted.