Response to the Editor

Dear Paul,

Thank you for your decision letter and for drawing attention to the interpretation of the grouped-by-Site cross-validation results. In our accompanying response to the referee, we provide a detailed description of how we clarified the model validation section in the manuscript.

At Trail Valley Creek, the landscape is strongly heterogeneous: some Sites (measurement locations) represent common combinations of vegetation, moisture and microtopography, whereas others capture rare or nearly unique conditions. When we perform grouped-by-Site cross-validation and hold out all data from a given Site, the environmental space occupied by that Site (for example, NDWI, TWI, landscape classification, microtopography) can be largely absent from the remaining training data. In such cases, the models are forced to extrapolate to combinations of predictors that have no close analogue in the training set, and it is therefore expected that R^2 values drop to ~0.1 - 0.2 in this very conservative scenario.

By contrast, the actual prediction task in this paper is to upscale CH_4 fluxes within the Trail Valley Creek area, across pixels that mostly fall inside the joint environmental space spanned by all chamber sites. For this within-domain setting, standard stratified k-fold cross-validation across individual measurements provides the relevant estimate of model performance, and those results show that RF and GBM capture spatial and temporal variability in CH_4 fluxes well. The grouped-by-Site analysis should therefore be interpreted as revealing the limits imposed by the small number of measurement locations in rare habitat types and the strong surface heterogeneity, rather than as evidence that the models are fundamentally unreliable.

We hope this clarifies why the grouped-by-Site results are not in conflict with our main conclusions and why the upscaling remains appropriate for the defined scope of the study, namely the heterogeneous surface at Trail Valley Creek.

To further align the body of the manuscript with the Abstract and highlight the utility of key remotely sensed variables, we have also added a sentence to the Abstract discussing the support offered by seasonal subsidence derived from remote sensing. This variable reflects important moisture gradients and shows high potential for improving CH_4 upscaling.

Sincerely,

Kseniia Ivanova

on behalf of all co-authors

Response to Referee

RC: 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.

AC: We apologise if our earlier reply created confusion about how cross-validation was implemented in the main analysis. Our previous wording may have suggested that the general model optimisation and all reported performance metrics were based on grouped-by-Site cross-validation, which is not the case.

In the main analysis, model tuning and performance assessment for the general models are based on standard stratified k-fold cross-validation across individual measurements, without grouping by Site or Year. All R^2, RMSE and MAE values reported in Tables 2 and 3 come from out-of-fold predictions of this standard k-fold CV. This setup corresponds to the prediction task we address in the paper: upscaling CH_4 fluxes within the Trail Valley Creek wetland complex, across pixels that mostly fall inside the joint environmental space spanned by all chamber sites.

By contrast, the grouped-by-Site and grouped-by-Year cross-validation runs shown in Text S4 and Fig. S7 are additional, deliberately conservative tests and were not used for model optimisation or for the main performance metrics. At Trail Valley Creek, Sites are measurement locations within heterogeneous area, and some Sites represent rare or nearly unique combinations of vegetation, moisture and microtopography. When we hold out all data from such a Site in grouped-by-Site CV, these conditions (for example, NDVI, NDWI, TWI) can be largely absent from the remaining training data. The models are then forced to extrapolate to combinations of predictors that have no close analogue in the training set, and in this strict setting it is expected that R^2 values drop to ~0.1–0.2.

We therefore interpret the grouped-by-Site CV results not as evidence that the general models are fundamentally unreliable, but as a stress test that reveals (i) how strongly performance deteriorates when entire measurement locations with rare combinations of conditions are removed, and (ii) how limited replication in rare habitat types constrains spatial transferability. For the within-domain upscaling task considered in this paper, the standard stratified k-fold cross-validation provides the relevant measure of model performance, and under this

evaluation RF and GBM reproduce the spatial and temporal variability in CH_4 fluxes reasonably well.

To avoid further ambiguity, we have clarified in Section 2.3.2 (“Model training and evaluation”), in the Results where we discuss Text S4 and Fig. S7, and in Supplementary Text S4, that (i) the main performance metrics are based on standard stratified k-fold CV across individual measurements, and (ii) the grouped-by-Site and grouped-by-Year CV are presented as additional diagnostic tests of transferability and sampling limitations rather than as the primary validation procedure.

Changes in text: Lines 290 – 293; added 3 lines after line 471

Minor comment :

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

AC: We have changed the capture accordingly.

RC: 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 ofmeasurements, 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.

AC:. We have changed Figure 4, and it’s the capture accordingly. Now it includes a number of sites as well.

RC: 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.

AC:. We have changed capture for Fig 7 accordingly.