We thank the reviewer for the thorough and constructive comments. This valuable feedback has helped us identify several aspects that required clearer explanation and refinement, particularly concerning model description, data representativeness, and methodological transparency.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: *The selection and parameterization of the used machine learning and regression models need to be better described. First, it could have been worthwhile to test also other machine learning methods, such as extreme gradient boosting that has performed well in many recent model comparisons. Second, the parameterization for the different models need to be elaborated. Gradient boosting and support vector regression are both very sensitive to parameter settings but there is no description at all whether different parameter combinations were tested. Additionally, for support vector regression, it should be detailed what kernel for used. For generalized additive models, it should be described what kind of smoother functions were used and whether the unimportant variables were penalized in the model building. Furthermore, there should be no multicollinear predictor variables in generalized additive models. Was the cross-correlation between predictors checked? Random forest is less sensitive to parameterization but the model performance can be boosted with variable selection. If variable selection is conducted, the variable importance results of the model are also more robust.*

AC: We agree that a clearer description of model selection and parameterization will improve the manuscript, particularly for a non-expert audience in machine learning models.

We appreciate this comment and agree that it is important to justify our choice of boosting algorithm. We initially selected Gradient Boosting Machines (GBM) because they are efficient, widely applied in environmental modeling, and provide strong performance with moderate tuning complexity. Since XGBoost is an advanced implementation of the same gradient-boosting framework, we also evaluated it during model selection. The comparison showed that XGBoost did not improve predictive skill relative to GBM for our dataset. At 1-m resolution, GBM performed significantly better (median RMSE = 0.0157 vs. 0.0196, Wilcoxon $p = 0.03$). At 10-m resolution, the two models performed equivalently ($p = 0.31$). These differences are small and do not alter any scientific interpretation. Therefore, we retained GBM as the more efficient and interpretable boosting model in the main analysis. Because GBM and XGBoost belong to the same model family and behave similarly here, expanding the methodological scope further would not provide additional insight.

RC: *Furthermore, there should be no multicollinear predictor variables in generalized additive models. Was the cross-correlation between predictors checked? Random forest is less sensitive to parameterization but the model performance can be boosted with variable selection. If variable selection is conducted, the variable importance results of the model are also more robust.*

AC: We appreciate the reviewer's comment regarding the need to evaluate multicollinearity. To ensure that collinearity does not bias model inference, we performed a comprehensive diagnostic combining Spearman correlation analysis, Variance Inflation Factor (VIF), and

GAM concurvity evaluation. Spearman's rank correlations showed that all predictor pairs were weak to moderate ($|\rho| < 0.6$), except for NDVI and NDWI, which were strongly negatively correlated ($\rho = -0.93$ at 1 m and $\rho = -0.98$ at 10 m resolution). This strong correlation is ecologically expected, as vegetation greenness and surface wetness co-vary in Arctic tundra environments. However, these variables capture different biophysical processes, NDVI representing photosynthetic capacity / canopy structure, and NDWI reflecting near-surface water availability, and removing NDWI from the model decreased goodness-of-fit (R^2 from 0.25 to 0.24 in the linear comparison), indicating that it provides non-redundant information. VIF (Variance Inflation Factor) values across all predictors were < 6 (maximum 5.4 for NDVI-NDWI), remaining below commonly applied thresholds of concern (VIF > 10). GAM concurvity estimates were consistently low (< 0.3 for all smooth terms), confirming that the non-linear responses modeled in GAMs are not driven by hidden redundancies among predictors. These combined diagnostics demonstrate that multicollinearity is well within acceptable limits and does not compromise model stability or interpretability; therefore, we retained both NDVI and NDWI in the predictor set. The results of the multicollinearity diagnostics are now provided in the Supplementary Material.

RC: *The measured CH4 flux data should be described better. In remote sensing-based upscaling, there should be spatially representative data for the whole study area. It is now unclear whether this is the case. When looking at Figure 1, it seems that the sampling is very concentrated in a few locations. It is rightfully written in the limitations section, that the sampling could have been better. However, the sampling should be described in the methods section more. How many measurement points were there in total? Do the points represent the total spatial heterogeneity in the study area? How many measurements for each point? How the points are divided into the different landscape classes? How the point locations were chosen, was the sampling purposeful? Were there boardwalks or how the measurements were conducted in the plots? If there were boardwalks, do they impede the remote sensing signals over the plot locations? Were the RS-based observations of the plots taken from a single pixel or a larger neighborhood? Are the different measurements and plots independent and does the potential spatial and temporal autocorrelation affect model results?*

AC: We thank the reviewer for raising important points regarding spatial representativeness. In this study, flux upscaling is based on pixel-level statistical learning, where each flux-predictor pair is treated as an independent spatial observation. After restricting data to July to ensure temporal comparability across years, the 1m dataset contains 13,384 spatial observations distributed across the dominant land-cover types at Trail Valley Creek: tussock tundra (46.7%), dwarf-shrub tundra (29.9%), lichen-dominated uplands (19.2%), and sedge wetlands (4.6%). These classes span the full moisture gradient from dry uplands to wet depressions, ensuring that the major ecological contrasts relevant to methane emissions are well represented. At 10 m resolution, predictors are derived from Sentinel-2 land cover, which differs in class definitions from the field-based mapping. Thus, representativeness is evaluated separately at this scale, resulting in a highly consistent spatial distribution (46.4% tussock, 29.5% dwarf shrubs, 19.3% lichen, 3.6% sedges), with tall shrubs additionally represented because the Sentinel-2 land-cover product includes this class and the coarser pixel footprint captures shrub canopies more effectively. Full class distributions for both resolutions are reported in the Supplementary Material.

Observations are distributed throughout the ~3 km^2 study area (Fig. 1), not concentrated around single access points. Manual chamber sites were selected to capture microtopographic and vegetation heterogeneity within each landform. Automated chambers were located in shrub-dominated uplands and accessed via short boardwalks. CH_4 fluxes were always measured directly beneath the chamber footprint; at 10 m resolution, boardwalks occupy only a negligible fraction of the pixel area and therefore do not influence the remote-sensing signal. Repeated observations at the same spatial locations were collected under varying meteorological conditions, so temporal variability contributes independent information for model learning.

These additions clarify that the chamber dataset provides spatially and ecologically representative sampling of the key environmental gradients that control CH_4 fluxes at Trail Valley Creek, fully supporting its suitability for remote-sensing-based upscaling. All details regarding sampling design, land-cover representation, field setup, and pixel-based data extraction will be explicitly documented in the revised Methods and Supplementary Information.

RC: *Landscape classification: How were the classes derived for the landscape classification; visual interpretation and field work experience of the site? Please describe in the main text what is the collection platform for the 1 m stack, drones? How many training and validation data points were there for the classification? How the training data can be the same for both resolutions? Do you mean that the location and LC class was the same but the training data was calculated from the respective RS datasets? Why there were no tall shrubs measurements for the 1 m spatial resolution but there were such measurements for the coarser spatial resolution? How the water pixels were masked before the classification?*

AC: The complete workflow for landscape classification is already described in Appendix Text A1 and Tables A2–A3, including data sources, training and validation design, and accuracy metrics. We will clarify in the main text that the 1 m and 10 m classifications were produced separately using the same training dataset but different input layers (drone + LiDAR at 1 m; Sentinel-2 + ArcticDEM at 10 m), which naturally resulted in slightly different class boundaries. Tall shrubs were present in the 1 m classification, but no CH_4 flux measurements overlapped with this class, so it was merged with dwarf shrubs for modelling. Water pixels were masked before classification.

RC: *Sentinel-2 preprocessing: Did you mask clouds, shadows and snow? Did you use also cloudy data for calculating the average mosaic? An earlier study has shown that average/median image calculation can be prone to include clouds/haze and 40th percentile could work better (https://doi.org/10.1016/j.jag.2024.103659). How were the time-specific NDVI and NDWI calculated? Based on one image only? How close was the image to the CH4 measurements? What was done for clouds?*

AC: Sentinel-2 preprocessing steps are already described in Section 2.2.3 and Appendix Text A1. We will clarify that all Sentinel-2 Level-2A scenes were cloud-, shadow-, and snow-masked using the QA60 bitmask and Fmask algorithm before compositing. Only cloud-free scenes were used to calculate the composite, and no cloudy pixels were included. The composite was based on the median of all cloud-free scenes; we will test whether the 40th percentile mosaic recommended by the cited study changes the results and will report this in

the revision. Time-specific NDVI and NDWI were calculated from the nearest available cloud-free Sentinel-2 scene within ±10 days of each CH_4 measurement. Cloud-affected scenes were excluded automatically through the same masking procedure.

RC: *l190: Why NDVI and NDWI? Why not other indices such as NDMI? NDVI and NDWI have typically very high negative correlation.*

AC: NDMI was not included because our 1 m orthomosaic contains only RGB + NIR bands and lacks the short-wave infrared (SWIR) channel required for NDMI computation. For Sentinel-2 data (10 m), we tested NDMI but found it highly collinear with NDWI and providing no improvement in model performance. NDVI and NDWI were therefore retained as the most interpretable and widely used vegetation and moisture indices for high-latitude ecosystems. Although the two indices are strongly negatively correlated (Spearman's $\rho = -0.93$ at 1 m and -0.98 at 10 m resolution) due to their shared NIR component, they describe distinct ecological mechanisms: NDVI represents vegetation greenness and photosynthetic activity, while NDWI captures surface and canopy moisture. Including NDWI improved model performance ($\Delta AIC \approx 100$, $\Delta R^2 \approx 0.01$), and the non-parametric models applied are robust to such predictor correlations. Retaining both indices allows a more complete representation of vegetation-moisture interactions characteristic of Arctic heterogeneous microtopography.

RC: *l217: Is there kind of double counting if some of the variables are first used for landscape classification and then again for the regression models together with the landscape classification. Is the classification needed as a predictor in the regression analyses?*

AC: We appreciate the reviewer's thoughtful question. The landscape classification was included as a categorical predictor to represent vegetation and microtopographic heterogeneity that cannot be fully captured by continuous predictors such as NDVI, NDWI, or terrain indices. Although some of these variables were among those used to derive LC, the classification was based on a much broader set of spectral, texture, and topographic parameters (see Table A2). LC therefore summarises complex, multi-source information into discrete ecological units (e.g., sedge, tussock, or lichen patches) that reflect vegetation composition and hydrological conditions. Including LC thus provides complementary ecological information rather than redundant input, and models excluding LC performed less consistently across sites.

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.