Supplement of Biogeosciences, 23, 233–262, 2026
https://doi.org/10.5194/bg-23-233-2026-supplement

Supplement of

High-resolution remote sensing and machine-learning-based upscaling of methane fluxes: a case study in the Western Canadian tundra

Kseniia Ivanova et al.

Correspondence to: Kseniia Ivanova (kivanova@bgc-jena.mpg.de)

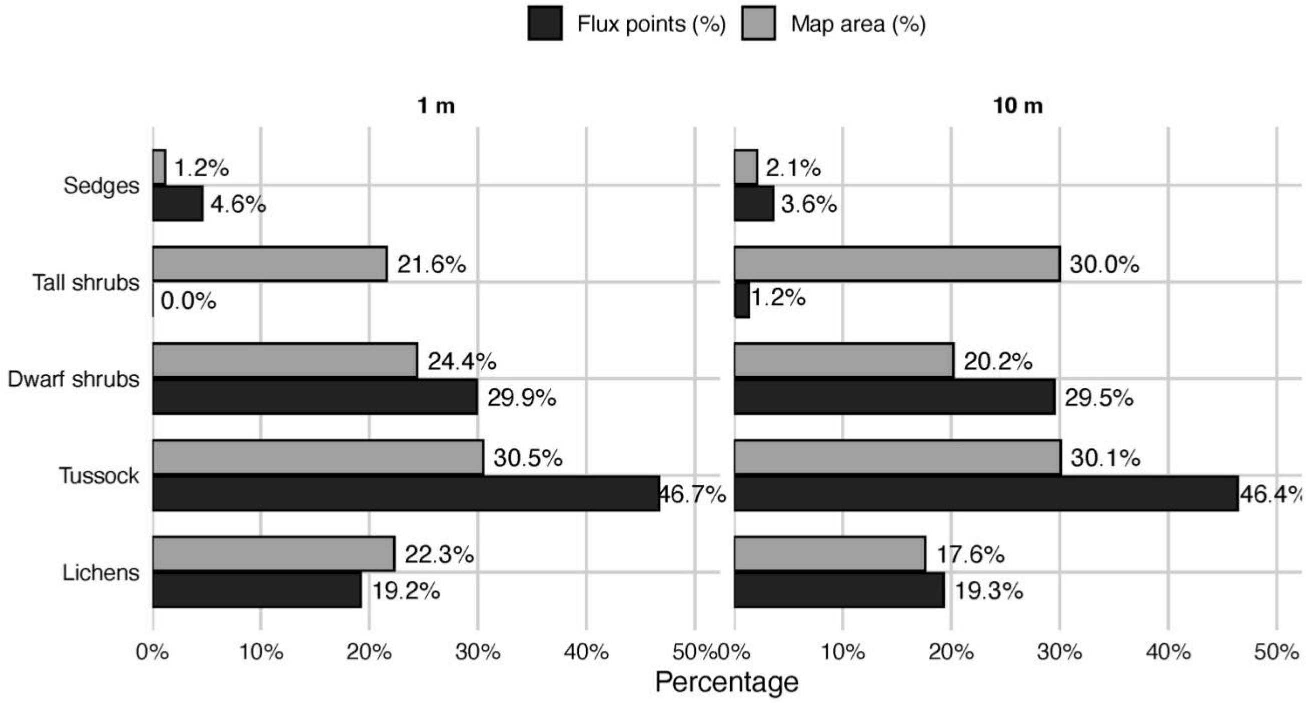


Figure S1. Spatial representativeness of chamber measurements across land-cover types at 1 m (left) and 10 m (right) resolution in Trail Valley Creek. Bars show the proportional map area (light grey) versus the proportion of flux observations (dark grey) for each land-cover class. Chamber sampling broadly reflects the dominant surface types in the AOI (tussock, dwarf shrubs, lichens), while sedges and tall shrubs occupy a small area and thus contribute fewer flux points. Values represent July flux observations only (n = 13,384).

Table S1. Class-specific accuracy for 1 m and 10 m landscape classifications. The validation set included 28 independent points (20 % of the total 140 training + validation samples).

Landscape class	**Test points (n)**	**Accuracy (1 m)**	**Accuracy (10 m)**
Lichen	6	67 %	67 %
Tussock	8	75 %	60 %
Dwarf shrub	2 (low n)	– (unstable)	– (unstable)
Tall shrub	10	80 %	89 %
Sedge	2 (low n)	– (unstable)	– (unstable)

S1. Hyperparameter settings and model configurations

Random Forest (ranger package)

The following hyperparameters were tuned:

- **mtry**
  Definition: number of predictors randomly sampled at each tree split.
  Effect: lower values increase diversity among trees; higher values strengthen each tree.
  Tested values: 5, 10

- **min.node.size**
  Definition: minimum number of samples in terminal nodes.
  Effect: smaller values capture fine-scale variability; larger values smooth predictions.
  Tested values: 5, 10

- **num.trees**
  Definition: number of trees in the forest.
  Effect: more trees increase stability but slow down computation.
  Fixed value: 500

Gradient Boosting Machine

- **n.trees**
  Definition: number of boosting iterations.
  Effect: higher values reduce bias but risk overfitting.
  Tested values: 500, 1000

- **interaction.depth**
  Definition: maximum depth of individual trees.
  Effect: determines flexibility and interaction order.
  Tested values: 3, 5

- **shrinkage**
  Definition: learning rate.
  Effect: lower values improve stability but increase runtime.
  Tested values: 0.01, 0.1

- **n.minobsinnode**
  Definition: minimum number of observations in terminal nodes.
  Effect: affects smoothness versus sensitivity to extremes.
  Tested values: 10, 20

Support Vector Regression

- **kernel**
  Definition: mapping of predictor space for nonlinear regression.
  Setting: Radial Basis Function (RBF), fixed
  Linear and polynomial kernels were also evaluated but showed substantially weaker predictive performance. Therefore, only the RBF kernel was retained in the final models.

- **sigma**
  Definition: kernel bandwidth.
  Effect: controls smoothness; low values allow sharp changes.

Tested values: 0.5, 1, 5

• **C**
Definition: penalty for model complexity.
Effect: high values reduce regularization and may overfit.
Tested values: 50, 100, 500

Generalized Additive Models

Generalized additive models (GAMs) were implemented using thin-plate regression splines (s() in mgcv) for all numeric predictors, while vegetation class was treated as a categorical term. Smoothing parameters were estimated with the REML method. We tested configurations with and without penalization of uninformative smooth terms (select = TRUE/FALSE) as well as two gamma settings (1.0 and 1.4) to control overfitting. The optimal configuration, selected based on lowest cross-validated RMSE, used select = TRUE and gamma = 1.4.

Predictor transformations

For RF and GBM we tested two formulations:

1. Linear

   All predictors entered directly in the model.

2. Polynomial

   Second-order polynomials applied to numeric predictors to test sensitivity to interaction structure.

SVR relies on kernel-based nonlinear mapping, and GAM uses spline-based nonlinear effects; therefore only linear predictor formulation was used for these models.

Final model hyperparameters

The optimal hyperparameter settings for each model were selected based on the lowest cross-validated RMSE and highest R^2. For Random Forest, the final configuration used mtry = 5, min.node.size = 5, splitrule = variance, and num.trees = 500. For Gradient Boosting Machines, the selected settings were n.trees = 1000, interaction.depth = 5, shrinkage = 0.1, and n.minobsinnode = 10. These tuned configurations were refit on the full dataset at each spatial resolution to generate the final CH_4 flux predictions.

S2. Collinearity diagnostics for GAM model predictors

Collinearity among predictors was evaluated using Variance Inflation Factors (VIF), which quantify the extent to which a predictor can be explained by other predictors in the model. VIF values > 5 typically indicate that a variable shares substantial information with another variable, and values > 10 are conventionally considered problematic. Because Landscape Class (LC) has several categories, we used a generalized form of VIF (GVIF) designed for categorical variables. We also adjusted the values so that all predictors can be compared on the same scale.

Table S2. Predictor collinearity check. Generalized Variance Inflation Factor (GVIF) results for the predictors used in the CH_4 flux models at 1 m and 10 m spatial resolution. The values show how strongly each predictor overlaps with the information provided by the other predictors. Values above 5 indicate moderate redundancy. Values above 10 indicate strong redundancy. Abbreviations: AT – air temperature; PAR – photosynthetically active radiation; TDD – cumulative thawing degree days; TPI – topographic position index; TWI – topographic wetness index; NDWI – normalized difference water index; NDVI – normalized difference vegetation index; Aspect – slope orientation; Slope – terrain steepness; LC – landscape class.

Predictor	1 m GVIF	10 m GVIF
AT	1.06	1.05
PAR	1.04	1.04
TDD	1.02	1.01
TPI	1.28	7.48
TWI	1.33	6.52
NDWI	2.48	7.80
NDVI	2.61	8.12
Aspect	1.21	5.32
Slope	1.45	4.73
LC	1.14	3.20

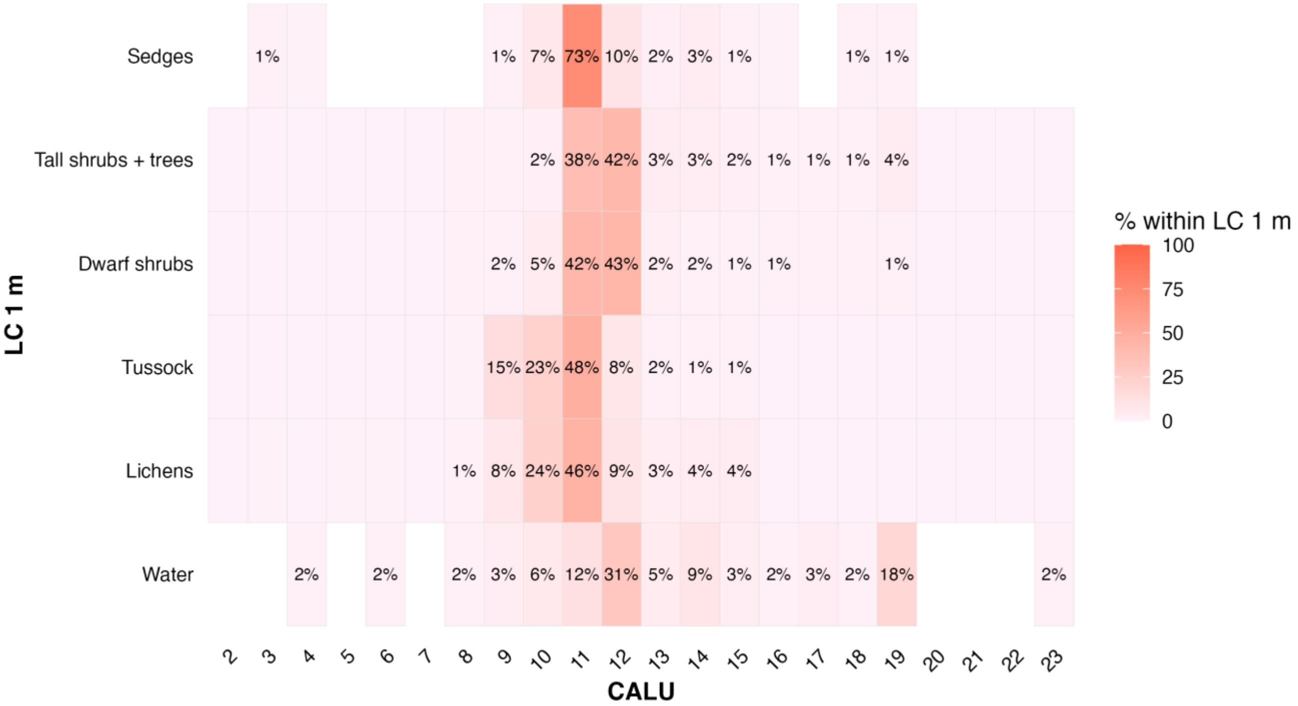
Collinearity diagnostics showed that the majority of predictors exhibited low to moderate redundancy at both spatial resolutions. Air temperature (AT), PAR, TDD, and LC had low adjusted VIF values (< 2), indicating that they contributed largely independent information to the models. Terrain-derived variables (Aspect, Slope, TPI, TWI) showed moderate redundancy, reflecting their shared geomorphological controls on drainage and microtopography.

In contrast, the satellite-derived moisture and vegetation indices (NDWI and NDVI) showed the highest VIF values, particularly at 10 m resolution. This pattern is expected, because spatial aggregation reduces fine-scale variation and strengthens correlations between moisture, vegetation, and topography. Despite this, their adjusted VIF values remained below widely used thresholds of concern (GVIF< 5), indicating that collinearity was not high enough to destabilize model fitting or inflate parameter uncertainty.

Generalized additive model (GAM) concurvity results confirmed that nonlinear smooth terms did not show problematic overlaps. Estimated concurvity values were low across predictors, meaning that each smoother explained a unique portion of CH_4 variability. This supports the use of spline terms without the need to remove predictors due to redundancy.

Together, these results demonstrate that all predictors can be retained in the final models without violating statistical assumptions related to multicollinearity. We therefore included the full predictor set in modeling at both 1 m and 10 m resolutions.

Figure S2. Pixel-wise cross-comparison between the CALU map (10 m resolution) and the site-specific landscape classification aggregated from 1 m to 10 m resolution. CALU represents a published pan-Arctic land-cover product, while LC 1 m originates from drone and LiDAR data classified at 1 m and then aggregated to 10 m by majority vote for comparability. Each cell in the matrix shows the percentage of pixels of a given CALU class that fall within a given LC class, so each CALU row sums to 100 %. Numbers are shown for values > 0.5 %. Coloured but unlabeled tiles indicate < 0.5 %, and blank tiles indicate that no such class combination occurs within the study area.



S3. Extension of the 10 m models with CALU and Subsidence predictors, and temporal NDVI/NDWI indices

In order to test whether large-scale land-cover products and temporal vegetation dynamics could further improve model performance, we performed an additional set of analyses using the 10 m dataset. Specifically, we included two broader-scale predictors: CALU (Circumarctic Land Cover Units; Bartsch et al., 2024) and Subsidence (InSAR-derived seasonal ground displacement), as well as temporally matched NDVI and NDWI values derived from Sentinel-2 imagery acquired within ± 10 days of chamber measurements. These analyses complement the main results (Section 3.4) by evaluating the potential contribution of geophysical deformation and short-term vegetation dynamics beyond the static and locally derived predictors used in the core models.

1. CALU and Subsidence

We extended the 10 m Random Forest (RF) and Gradient Boosting Machine (GBM) models by adding CALU and Subsidence as predictors. These variables capture broader environmental context: CALU representing vegetation composition and surface type derived from pan-Arctic classification, and Subsidence reflecting surface deformation and seasonal thaw-related ground movement.

Subsidence emerged as the second most influential predictor (Fig. S3) in both models, underscoring its strong association with soil moisture dynamics and active-layer processes that directly influence CH_4 fluxes.

CALU contributed moderately in RF (≈ 2.5 %) but more strongly in GBM (≈ 15 %), where it exceeded the locally derived landscape class (≈ 1 %). This reversal (CALU > LC in GBM vs. LC > CALU in RF) reflects algorithm-specific sensitivities: RF emphasises fine-scale categorical heterogeneity, whereas GBM integrates additive patterns related to vegetation and moisture gradients.

2. Temporal NDVI and NDWI

To evaluate whether temporally matched vegetation and wetness indices improve CH_4 flux prediction, we replaced the static NDVI and NDWI layers with Sentinel-2 indices closest in time to the chamber measurements (Fig. S4). This modification substantially altered the predictor hierarchy at 10 m resolution. In the RF model, topography-related variables gained influence, with TWI emerging as the strongest driver, followed by Slope and landscape class. In contrast, NDVI and NDWI lost their previous prominence, suggesting that temporal variability in these indices introduced additional noise rather than new explanatory signal. The GBM model showed a similar shift: while TWI remained the leading predictor, Air Temperature became comparably influential, and NDWI retained moderate importance. Overall, temporally dynamic NDVI and NDWI did not improve model interpretability or dominance among predictors, indicating that for short Arctic growing-season windows, static indices capture vegetation and surface wetness patterns more robustly than temporally matched scenes affected by cloud-related gaps and sensor timing differences.

Figure S3. Relative importance (%) of environmental predictors for CH_4 fluxes in the 10 m models including Subsidence and CALU (Circumarctic Land-cover Units) as additional predictors. Results are shown for Gradient Boosting Machine (GBM) and Random Forest (RF) models. Importance was estimated using permutation-based resampling and normalised within each model.

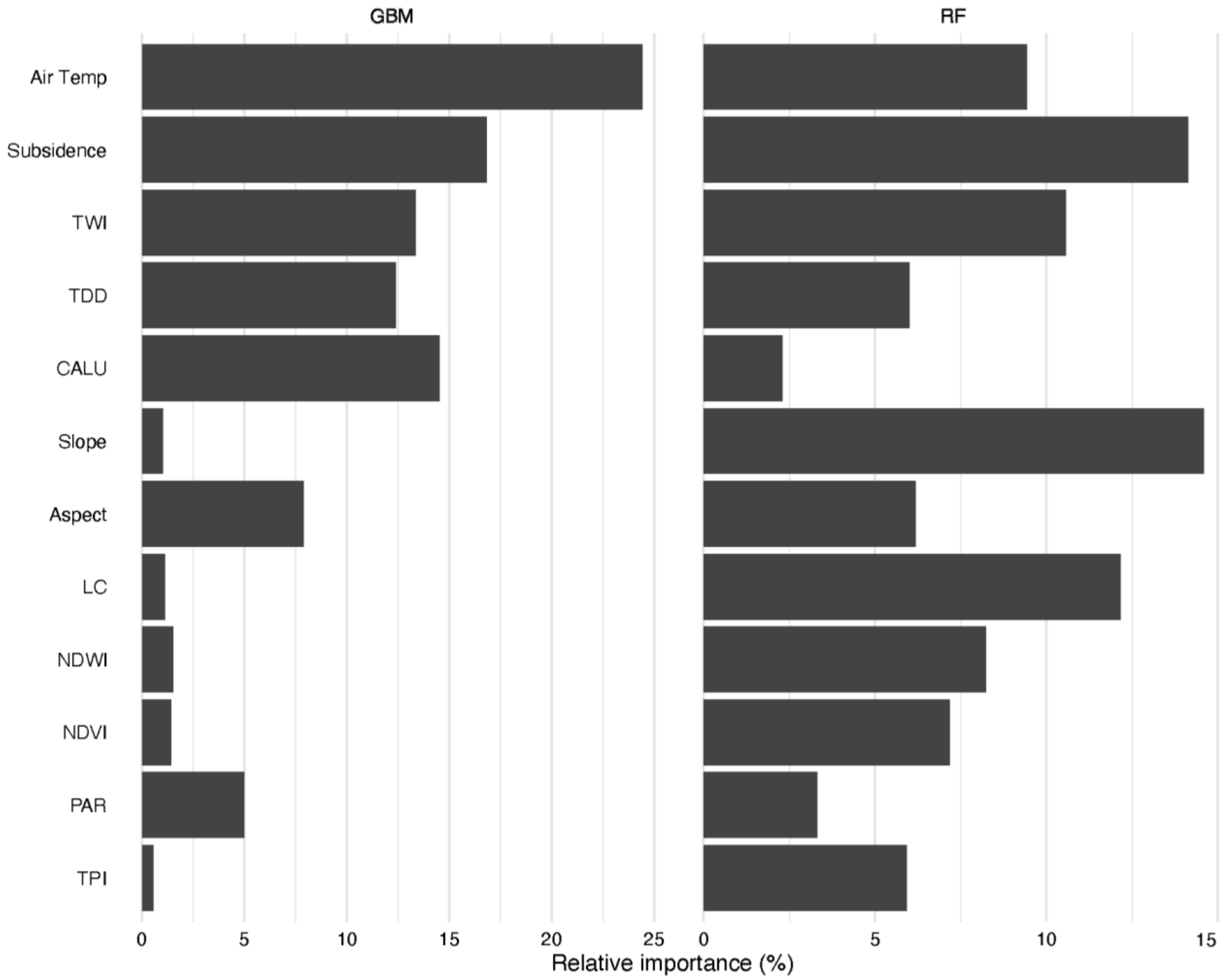
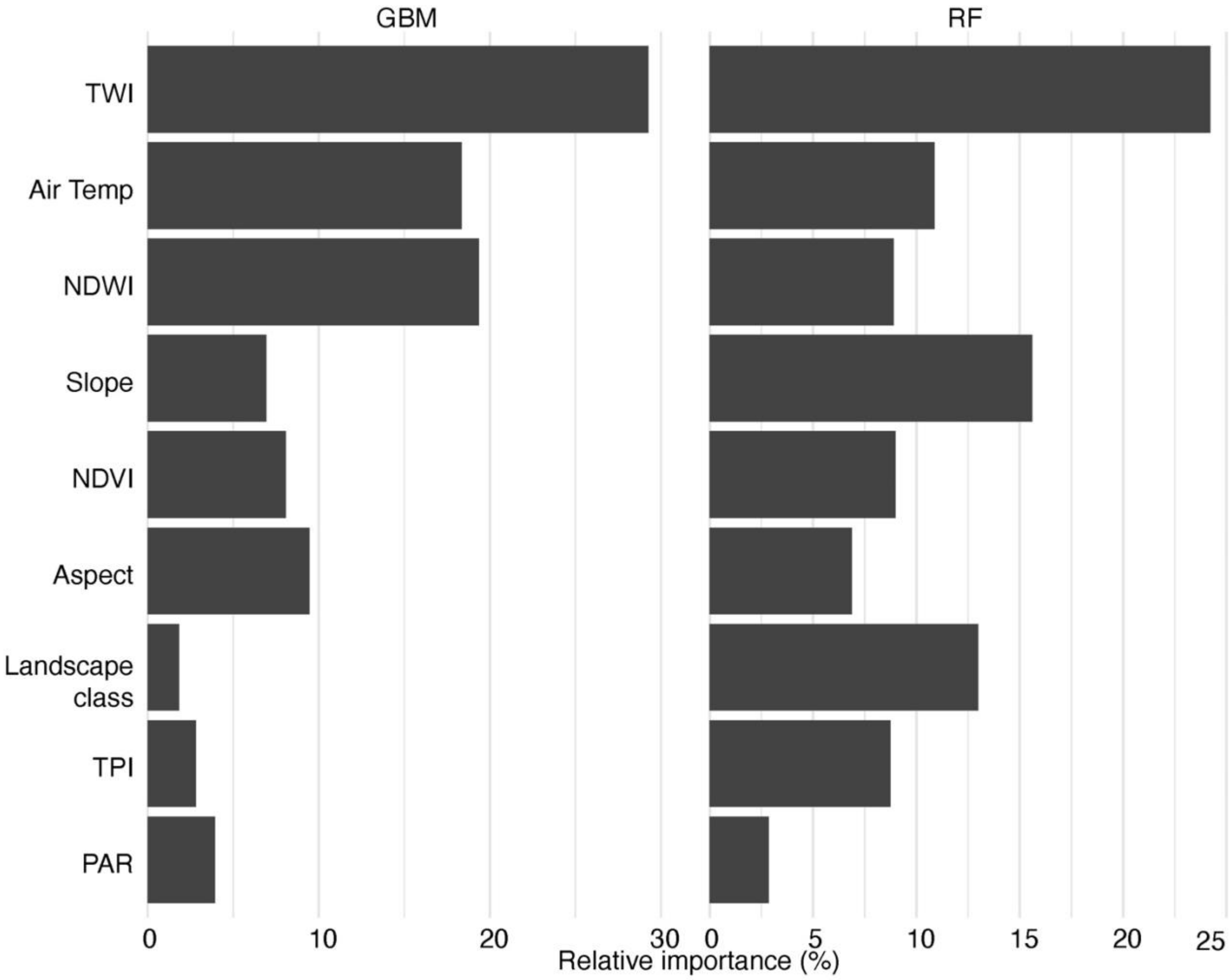


Figure S4. Relative importance of environmental predictors in 10 m Random Forest (RF) and Gradient Boosting Machine (GBM) models using temporally matched NDVI and NDWI.



S4. Resolution aggregation test to separate data source and scale effects

The comparison between 1 m and 10 m model performances inherently combines two sources of variation:

1. Differences related to spatial resolution (e.g. averaging across larger grid cells)
2. Differences between the underlying datasets (e.g. acquisition time, sensor characteristics).

To isolate these influences, we created an additional dataset where the 1 m input layers were aggregated to 10 m resolution using identical processing steps. This allowed us to isolate the effect of resolution from that of data source differences.

Numeric predictors (e.g. NDVI, NDWI, TPI, TWI, Slope, Aspect) were aggregated by mean, while categorical variables (LC) were aggregated by majority vote based on the dominant pixel class within each 10 m grid cell. The resulting "10 m from 1 m" dataset preserved the spectral and structural characteristics of the original high-resolution inputs but matched the coarser grid of the 10 m Sentinel-based data.

We compared the distributions of all predictors across the three datasets (1 m, 10 m, and 10 m from 1 m). The aggregated data were more similar to the original 1 m values (even when statistically different in most cases) than to the 10 m satellite-derived predictors (Fig. S5). For example, NDVI, Slope, and TWI retained the characteristic variability of the 1 m data, while the 10 m Sentinel-based inputs appeared smoother and less variable.

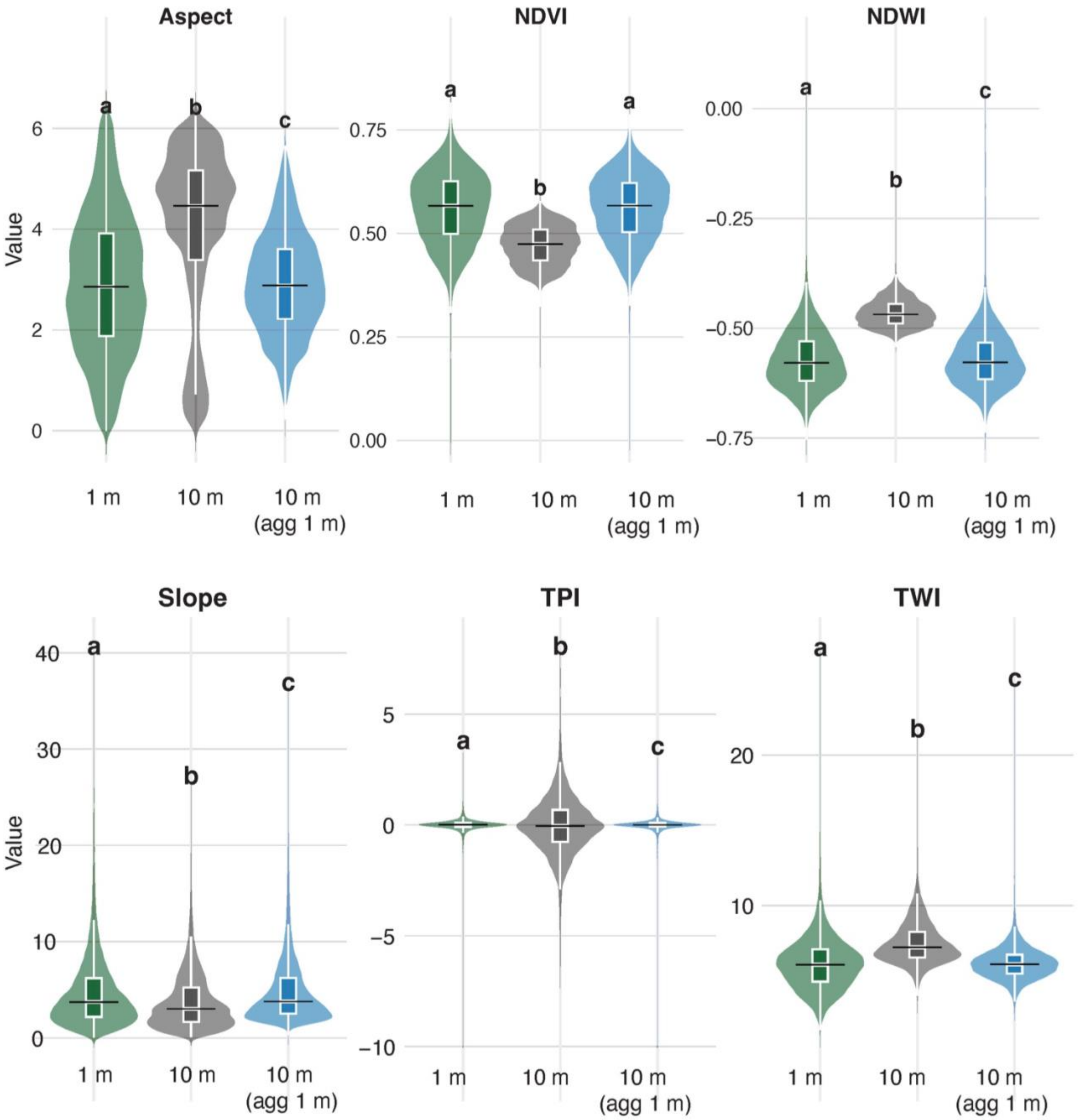


Figure S5. Comparison of predictor distributions across resolutions (1 m, 10 m, and 10 m aggregated from 1 m). Violin plots show the value range and median for each variable. Letters above violins indicate significant differences between groups (Tukey's HSD, $p < 0.05$).

To quantify the impact on model performance, we applied the same modelling workflow to the 10 m from 1 m dataset and compared results to the original 10 m models. Differences in RMSE were minor (within ±5 % across all algorithms): Random Forest = +4.1 %, GBM = −4.9 %, SVR = −4.3 %, GAM = +0.07 % (Fig. S6).

Overall, model performance at 10 m from aggregated 1 m data was very similar to that of the original 10 m models, with only minor differences across algorithms. Some models (e.g., GBM and SVR) showed slightly better accuracy after aggregation, while others (e.g., RF) performed slightly worse, and GAM remained nearly unchanged. These small differences suggest that the performance gap between the 1 m and 10 m models reported in the main text mainly reflects the effect of spatial resolution rather than differences in input data sources.

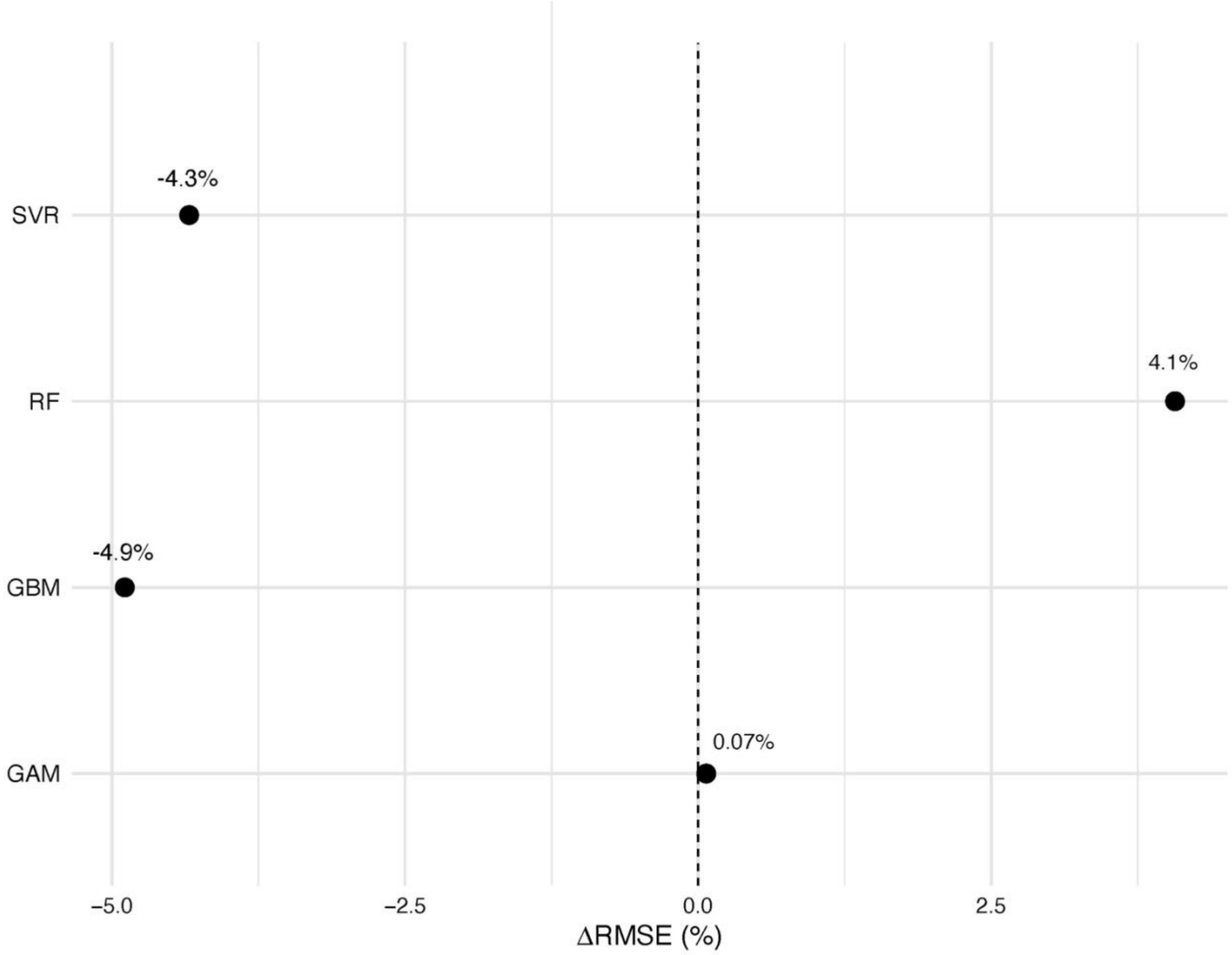


Figure S6. Comparison of model performance between 10 m and 10 m (from 1 m aggregated) datasets. Points show percentage differences in RMSE (ΔRMSE %) for four model types. Negative values indicate lower error in the aggregated dataset.

S5. Cross-validation sensitivity: grouped by site and by year

To evaluate how sampling structure affects model transferability, we repeated model evaluation using grouped cross-validation, where all data from the same year or site were held together in either the training or testing subsets. This setup avoids data leakage across correlated measurements and provides two complementary tests: (i) Year-CV, assessing temporal transfer to new measurement years, and (ii) Site-CV, assessing spatial transfer to unmeasured locations.

In this dataset, the two grouping factors are not independent. Measurement years correspond to partially distinct sets of sites and environmental conditions: automatic chambers, placed in relatively drier areas, were sampled in 2019 and 2021, whereas wetter areas were sampled using mobile chambers in 2022-2024. As a result, holding out an entire year or a set of sites often removes entire parts of the hydrological and vegetation gradients (NDWI, TWI, LC), forcing the model to extrapolate beyond its training distribution rather than interpolate within it.

This design imbalance explains the sharp decline in predictive performance when grouped CV is applied (Fig. S7). Under a standard five-fold CV, models achieve high accuracy (R^2 = 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 R^2 values (typically 0.1-0.2) and larger RMSE (0.15-0.4). These values do not indicate model instability but rather reveal that flux-environment relationships learned from one subset of the landscape cannot be directly transferred to sites or years that represent different ecosystem types.

Such behaviour is expected when predictor distributions differ strongly between training and test subsets. In practical terms, the grouped CV simulates a scenario of out-of-distribution prediction, for example, applying a model trained on moist sedge areas to dry dwarf-shrub areas, or vice versa. The observed decrease in R^2 therefore reflects the intrinsic spatial and temporal heterogeneity of the study area rather than model overfitting.

Overall, this sensitivity test demonstrates that the RF and GBM models are internally consistent within the sampled environmental space but cannot fully generalise to conditions not represented in the training data. These results highlight the need for more temporally repeated measurements at identical microtopographic locations to better quantify interannual predictability of site-level CH_4 dynamics.

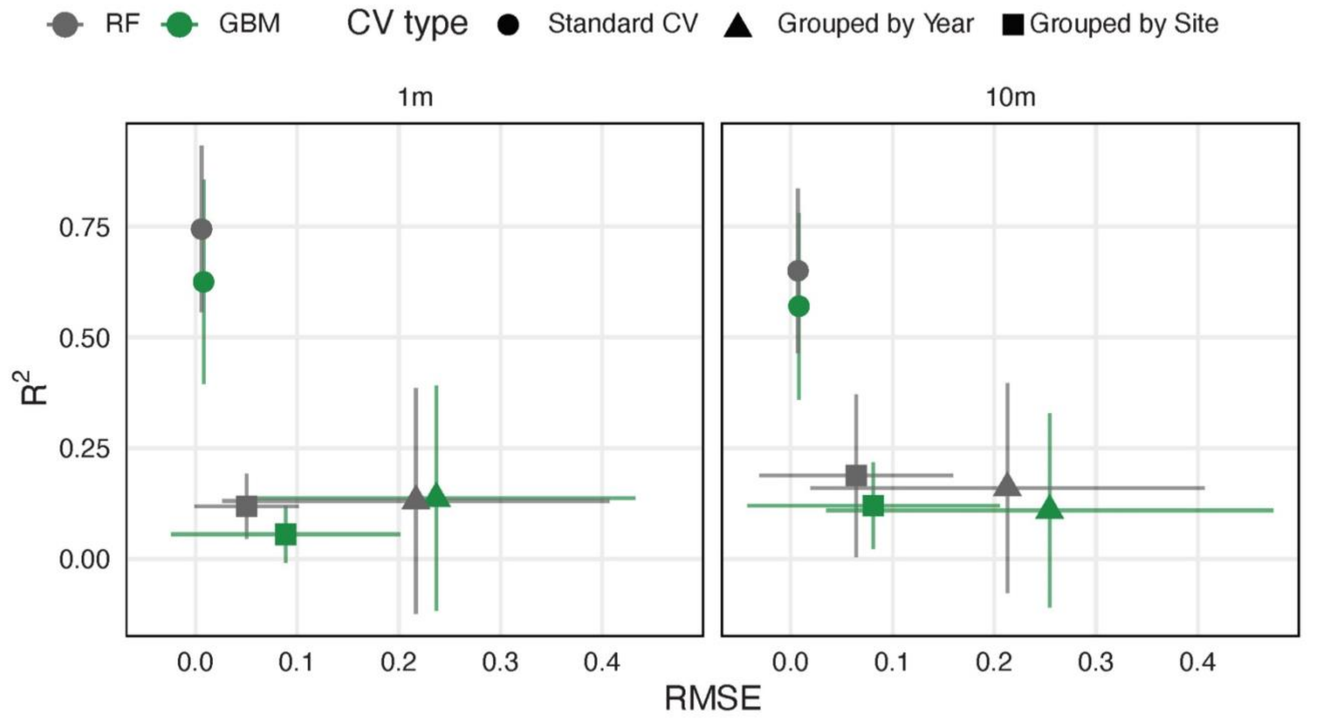


Figure S7. Model performance (R^2 vs. RMSE) for Random Forest (gray) and Gradient Boosting Machine (green) models under different cross-validation schemes (Standard CV = squares, Grouped by Year = triangles, Grouped by Site = circles) at 1 m and 10 m resolutions. Points show mean performance across folds;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