The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B3-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission III, 4–11 July 2026, Toronto, Canada

High-Resolution GPP Estimation from Sentinel-1 and Sentinel-2 Around AmeriFlux Towers: Evaluating Temporal and Geographic Generalisation

Taewoong Ham[1], Baoxin Hu[1]

[1] Department of Earth and Space Science and Engineering, York University, Toronto, Canada
twham21@yorku.ca, baoxin@yorku.ca

Keywords: Gross primary production, satellite imagery, deep learning, data fusion, generalisation, remote sensing

Abstract

Accurate estimation of gross primary production (GPP) is fundamental for quantifying the terrestrial carbon cycle. However, coarse-resolution products often fail to capture fine-scale spatial variations in carbon uptake across heterogeneous landscapes. While recent studies have begun to employ 10 m Sentinel-1 and Sentinel-2 imagery, they typically reduce these data to pixel-wise spectral indices, discarding the two-dimensional spatial structure (canopy architecture, land-cover transitions, within-stand heterogeneity) that the imagery encodes. This study investigates whether explicitly exploiting this spatial context via convolutional neural networks yields robust, transferable gains over tabular machine-learning baselines. We curate a quality-controlled dataset of 23,528 eight-day multi-sensor composites from 222 AmeriFlux sites (2015–2025), evaluated under site-wise cross-validation, temporal generalisation, and geographic transfer to an 18-site upper Midwest forest holdout. Under temporal transfer to unseen years (2023–2025), the best convolutional model achieves $R^2 = 0.77$ and RMSE $= 1.95\,\mathrm{gC\,m^{-2}\,d^{-1}}$, an 18.6% RMSE reduction over ridge regression ($R^2 = 0.65$, RMSE $= 2.40\,\mathrm{gC\,m^{-2}\,d^{-1}}$). Although this advantage narrows under geographic transfer to structurally novel regions ($R^2 = 0.59$ vs. 0.54), the convolutional models still outperform all tabular baselines. Spatial structure at 10 m therefore supports more robust temporal generalisation than spectral aggregates alone.

1. Introduction

Satellite-derived estimates of gross primary production (GPP) are widely used to monitor terrestrial carbon uptake. However, many commonly used products remain too coarse to resolve the fine-scale heterogeneity of real landscapes. The MODIS MOD17 product provides 8-day composites at 500 m resolution (LAADS DAAC, 2015), and FLUXCOM products operate at 0.0833° to 0.5° (Jung et al., 2020). Other light-use-efficiency products are often similarly coarse. For example, a global 0.05° dataset was provided by (Bi et al., 2022), while recent reviews summarise many coarse-resolution light-use-efficiency products (Pei et al., 2022, Zhu et al., 2024). At these scales, a single pixel can contain forests, croplands, wetlands, and open ground at the same time. Fine-scale variation in carbon uptake associated with canopy structure, species composition, management, topography, and disturbance is therefore averaged out.

This limitation has become more apparent with the availability of high-resolution Sentinel data. Sentinel-2 provides optical observations at 10 m resolution (Drusch et al., 2012), and Sentinel-1 provides radar observations that add structural information not available from optical data alone (Torres et al., 2012). Higher-resolution satellite data also reduce the mismatch between image pixels and flux-tower footprints (Kong et al., 2022). Sentinel-based GPP estimation has already been explored in radiative-transfer-trained models, flux-tower upscaling, and other high-resolution studies (Wolanin et al., 2019, Cai et al., 2021, Pabon-Moreno et al., 2022, Spinosa et al., 2023, Junttila et al., 2023, Vira et al., 2025). Yet most existing approaches still convert each pixel or patch into spectral values, vegetation indices, light-use-efficiency terms, or other summary features. The spatial arrangement of those values is therefore discarded, even though 10 m imagery contains information on within-stand heterogeneity, canopy edges, and land-cover transitions. Optical and radar measurements are also complementary, because they respond to different aspects of vegetation condition and structure (Joshi et al., 2016).

Convolutional neural networks can use this spatial information directly. Spatial structure often changes more slowly than the year-to-year atmospheric and phenological variability that affects spectral response, so models that learn from spatial patterns may transfer more robustly across time than models based only on patch-level summary features. The key test is whether that advantage persists in future years and in geographically distinct regions.

Tabular machine-learning baselines and convolutional neural networks with three fusion strategies are compared using 10 m Sentinel-1 and Sentinel-2 patches centred on AmeriFlux towers (Baldocchi et al., 2024). Performance is evaluated under site-wise cross-validation, temporal transfer to held-out years, and geographic transfer to an 18-site upper Midwest forest holdout. The comparison tests both in-distribution accuracy and robustness under the temporal and regional shifts relevant to practical application. The main contributions are:

1. a quality-controlled dataset of 23,528 8-day Sentinel-1 and Sentinel-2 patches paired with tower-derived GPP from 222 AmeriFlux sites (2015–2025), including an 18-site upper Midwest forest holdout for out-of-distribution evaluation;
2. a unified evaluation design spanning site-wise cross-validation, future-year transfer, and geographic transfer, showing that cross-validation alone overstates model robustness;
3. empirical evidence that convolutional models using two-dimensional spatial structure outperform tabular baselines, with the largest gain under temporal transfer (18.6% lower RMSE than ridge regression) and smaller but consistent gains under geographic transfer.

2. Dataset and Preprocessing

The target variable is 8-day mean GPP derived from AmeriFlux eddy-covariance observations (AmeriFlux, 2021). Daily values are drawn from the FLUXNET FULLSET product where available (Pastorello et al., 2020). The nighttime-partitioned variable `GPP_NT_VUT_REF` is used (FLUXNET, n.d.). The daily series are aggregated to MODIS-aligned 8-day periods, yielding 46 periods per year. Negative GPP values are set to zero, and only periods with at least 75% valid daily observations are retained. In the final dataset, GPP ranges from 0 to 31.69 $\mathrm{gC\,m^{-2}\,d^{-1}}$, with a mean of 3.16 and a standard deviation of 3.92.

Each sample is a 128×128 patch centred on a flux tower at 10 m resolution, corresponding to $1.28\,\mathrm{km} \times 1.28\,\mathrm{km}$ ($1.64\,\mathrm{km}^2$). Monthly 80% footprint climatologies across 214 AmeriFlux sites range from 10^3 to $10^7\,\mathrm{m}^2$ (Chu et al., 2021). The selected patch size therefore covers the tower footprint at many sites while retaining surrounding landscape context.

The optical input is derived from ten Sentinel-2 bands spanning the visible, red-edge, near-infrared, and shortwave-infrared regions (Drusch et al., 2012). Sentinel-2 Level-2A surface reflectance is used, and the clearest available observation within each 8-day period is selected at each pixel using the scene classification layer (Copernicus Sentinel Wiki, n.d.). Bands acquired natively at 20 m (B05–B07, B8A, B11, B12) are resampled to 10 m with bilinear interpolation. Where no clear observation is available, the temporal median reflectance is used instead. Reflectance values are divided by 10 000 and clipped to $[0, 1.2]$; the upper bound allows for bright targets whose apparent reflectance can exceed unity after atmospheric correction.

The radar input is derived from Sentinel-1 VV and VH backscatter (Torres et al., 2012). A radiometrically terrain-corrected product is used (Alaska Satellite Facility, n.d.). For each 8-day period, backscatter is composited as the temporal median in linear power and then converted to decibels. Values are mapped linearly from $[-30, +5]$ dB to $[0, 1]$. Periods with single-polarisation Sentinel-1 acquisitions are excluded. These fixed sensor-based normalisations avoid dependence on training-set statistics. Any remaining missing pixels after compositing are filled with the spatial median of the corresponding channel. A representative composite is shown in Figure 1.

Four additional channels provide temporal and spatial context. Two encode day of year with sin and cos transforms. Two encode patch-relative position as row and column coordinates scaled to $[0, 1]$, analogous to CoordConv-style coordinate channels (Liu et al., 2018). The resulting input tensor contains 16 channels in total (Table 1).

These preprocessing steps are applied uniformly to all candidate composites, but only a subset passes the full supervision and quality-control pipeline. Composite metadata initially cover 138,322 entries across 727 sites. Intersecting these records with the GPP label set yields 25,738 composite–label matches across 225 sites. Filtering on composite completeness, band availability, and GPP coverage retains 23,528 samples from 222 sites spanning 2015–2025 (Table 1). Of these, 22,475 samples come from 206 US sites and 1,053 samples come from 16 Canadian sites.

Geographic transfer is evaluated with a dedicated holdout from a spatially coherent cluster of 18 upper Midwest US forest sites.

Table 1. Dataset and input summary.

Statistic	Value
Total curated samples	23,528
Total sites (US / Canada)	222 (206 / 16)
Observed site-year pairs	1006
Year range	2015–2025
Geographic holdout sites / samples	18 / 923
Training pool (holdout excluded)	22,605 samples, 204 sites
Patch size	128×128 (1.28 km)
Input channels	16
Sentinel-2 (optical)	10
Sentinel-1 (synthetic aperture radar, SAR)	2
Day-of-year encoding	2
Spatial coordinates	2

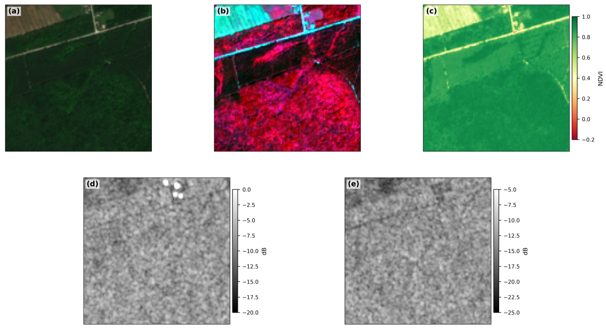

Figure 1. Example 8-day multi-sensor composite centred on an AmeriFlux tower: (a) Sentinel-2 true colour, (b) Sentinel-2 false colour, (c) normalised difference vegetation index (NDVI), (d) Sentinel-1 VV backscatter, and (e) Sentinel-1 VH backscatter.

This holdout contains 923 samples spanning 2016–2025. After excluding these sites, the remaining training pool contains 22,605 samples from 204 sites (188 US and 16 Canadian). The holdout sites are excluded from all training and validation folds.

3. Methods and Evaluation Design

All models (four tabular baselines and three convolutional architectures) are evaluated under identical site-wise cross-validation, temporal generalisation, and geographic holdout protocols.

3.1 Tabular baselines

The tabular baselines reduce each 128×128 pixel patch to a fixed-length feature vector by computing the spatial mean and standard deviation of each input band, yielding 24 spectral statistics (10 Sentinel-2 bands and 2 Sentinel-1 bands, each summarised by two statistics). The normalised difference vegetation index (NDVI, computed from B04 and B08) and the Sentinel-1 cross-polarisation ratio ($\mathrm{VH} - \mathrm{VV}$ in decibels) are derived and similarly summarised, contributing 4 additional features. Cyclical day-of-year encoding (sin and cos) contributes 2 additional scalar features, and the site land-cover class is one-hot encoded (eight categories across the dataset).

Four models are fitted to these features: ordinary least-squares linear regression, ridge regression ($\alpha = 1.0$), random forest (300 trees, unlimited depth), and gradient-boosted regression trees (GBRT; 500 trees, learning rate 0.05, maximum depth 3). Missing values are imputed with the per-feature median. Standard scaling is applied to the two linear models; the tree-based models operate on unscaled features.

3.2 Convolutional fusion models

Three convolutional architectures are compared, each operating on the full $128 \times 128 \times 16$ input tensor. All backbones are based on ResNet-18 (He et al., 2016) and are trained from scratch with Kaiming uniform initialisation (He et al., 2015) (no ImageNet pre-training, as the 16-channel multi-sensor input is incompatible with three-channel natural-image weights). Each model terminates in a regression head consisting of global average pooling, one or more fully connected layers with ReLU activation and 20% dropout, and a single linear output.

Early Fusion. All 16 input channels are concatenated and processed by a ResNet-18 backbone whose first convolutional layer is adapted to accept 16 rather than 3 input channels. The backbone produces a 512-dimensional feature vector that is mapped to a scalar GPP estimate through two fully connected layers ($512 \rightarrow 128 \rightarrow 1$; 11.3 M parameters).

Dual-Branch Late Fusion. The optical (10 Sentinel-2 bands) and radar (2 Sentinel-1 bands) streams are routed to independent encoders. The optical branch follows a full ResNet-18 architecture and produces a 512-dimensional representation; a lighter encoder with half the channel width and single-block residual layers processes the radar stream and produces a 256-dimensional representation. The four context channels are reduced to a 4-dimensional vector by global average pooling. The three representations are concatenated (772 dimensions) and passed through a fusion layer ($772 \rightarrow 128$) followed by the regression head ($128 \rightarrow 128 \rightarrow 1$; 12.5 M parameters).

Middle Fusion. Optical and radar streams each pass through a separate convolutional stem and one residual block, producing 64- and 32-channel feature maps at 32×32 spatial resolution. These maps are concatenated along the channel axis and projected to 64 channels by a 1×1 convolution, then processed by a shared three-stage residual backbone. Context channels are pooled and concatenated with the 512-dimensional backbone output (516 dimensions), which is passed through a fusion layer ($516 \rightarrow 128$) and the regression head ($128 \rightarrow 128 \rightarrow 1$; 11.2 M parameters). By fusing at the feature-map level rather than after global pooling, this architecture preserves spatial correspondences between the optical and radar modalities through the deeper layers.

3.3 Training procedure

All convolutional models are trained with the AdamW optimiser ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-4}) using an initial learning rate of 2×10^{-3} and a cosine annealing schedule decaying to 2×10^{-6} over 80 epochs. The loss function is the Huber loss with $\delta = 1.0$, chosen to down-weight large residuals that arise from eddy-covariance measurement noise and gap-filling artefacts in the GPP labels. Training uses half-precision floating-point (FP16) mixed-precision arithmetic, a batch size of 256, and gradient-norm clipping at 1.0. Early stopping halts training when the validation RMSE does not improve for 15 consecutive epochs, and the checkpoint with the lowest validation RMSE is retained.

Data augmentation is restricted to centre-preserving transforms so that the central pixel, which corresponds to the flux-tower location and thus the GPP label, remains at a fixed spatial position after transformation. The augmentation pipeline applies random horizontal and vertical flips (each with probability 0.5) and 90° rotations (probability 0.5). Modal dropout (probability 0.1) selects one of the two modalities uniformly at random and zeroes it, encouraging the network to learn from each modality independently. Additive Gaussian noise (probability 0.2, $\sigma \approx 0.004$ in normalised units, corresponding to 0.15 dB before normalisation) is injected into the radar channels.

3.4 Evaluation protocol

Model performance is assessed with two complementary metrics. The root mean square error,

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}\,, \tag{1}$$

measures the average prediction error in the original units ($\mathrm{gC\,m^{-2}\,d^{-1}}$); lower values indicate more accurate predictions. The coefficient of determination,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2}\,, \tag{2}$$

quantifies the proportion of variance in the observed GPP explained by the model, where $\bar{y}$ denotes the mean of the observed values; a value of 1 indicates perfect prediction and a value of 0 corresponds to predicting the mean. Three evaluation settings are reported:

1. **Cross-validation.** A site-wise five-fold split groups all samples from a given tower into the same fold, preventing data leakage between sites. Folds are assigned by a stratified greedy algorithm that balances five land-cover meta-groups across folds. Early stopping selects the checkpoint with the lowest RMSE on the held-out fold.
2. **Temporal generalisation.** All models are trained exclusively on 2015–2021 data and validated on 2022 for model selection; final evaluation is performed on a separate 2023–2025 test partition not used during training or model selection. The same site may appear in both training and test partitions in different years; this protocol therefore tests future-year prediction at previously observed sites, not transfer to entirely unseen locations.
3. **Geographic holdout.** Each of the five cross-validation models is evaluated independently on the 18-site upper Midwest forest holdout set, which is excluded from all training and validation folds. Results are reported as the mean and standard deviation across folds. Tabular baselines are retrained on each fold's training data and evaluated on the holdout in the same manner.

All reported results are from a single-seed experiment (seed 42).

4. Results

Table 2 summarises the performance of all seven models across the three evaluation settings. Cross-validation RMSE values range from 1.89 to 2.23 $\mathrm{gC\,m^{-2}\,d^{-1}}$; tabular and convolutional approaches achieve broadly comparable accuracy when evaluation sites are drawn from the same network and geographic distribution as the training sites. The gap between the two model families widens under temporal generalisation and persists, though attenuated, under geographic transfer.

Table 2. Performance of tabular baselines and convolutional fusion models under cross-validation (CV), temporal generalisation (Temporal), and geographic holdout (OOD) evaluation. RMSE is in $\mathrm{gC\,m^{-2}\,d^{-1}}$; higher R^2 indicates better variance explanation. OOD results report the mean ± standard deviation across five cross-validation folds. Bold marks the best value per column within each model group. All results are from a single seed.

	CV		Temporal		OOD	
Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
Linear	2.2124	0.6615	2.4022	0.6477	**3.3346 ± 0.02**	**0.5371 ± 0.005**
Ridge	**2.2110**	**0.6622**	**2.3976**	**0.6490**	3.3348 ± 0.02	0.5370 ± 0.005
Random Forest	2.2125	0.6600	2.8319	0.5103	3.52 ± 0.13	0.484 ± 0.037
GBRT	2.2299	0.6560	2.8999	0.4865	3.57 ± 0.09	0.468 ± 0.026
Early Fusion	1.8960	0.7494	1.9630	0.7647	**3.15 ± 0.05**	**0.587 ± 0.014**
Dual-Branch	**1.8909**	**0.7504**	1.9838	0.7597	3.26 ± 0.08	0.558 ± 0.021
Middle Fusion	1.9036	0.7459	**1.9517**	**0.7674**	3.29 ± 0.12	0.548 ± 0.032

4.1 Cross-validation

Within the site-wise five-fold cross-validation, all seven models produce RMSE values between 1.89 and 2.23 $\mathrm{gC\,m^{-2}\,d^{-1}}$, with corresponding R^2 values ranging from 0.656 (GBRT) to 0.750 (Dual-Branch). The three convolutional architectures cluster tightly between 1.89 and 1.90 in RMSE and 0.746–0.750 in R^2, while the four tabular baselines fall between 2.21 and 2.23 in RMSE and 0.656–0.662 in R^2. Relative to a dataset standard deviation of 3.92 $\mathrm{gC\,m^{-2}\,d^{-1}}$, the difference between the best convolutional model (Dual-Branch, 1.8909) and the best tabular model (Ridge, 2.2110) amounts to a 14.5% reduction in RMSE. When training and evaluation sites share the same geographic distribution, the spatial context that the convolutional models exploit provides only a limited advantage over patch-level spectral descriptors.

4.2 Temporal generalisation

A clearer separation between the two model families emerges under temporal generalisation, where models trained on 2015–2021 data are evaluated on 2023–2025. Among the tabular baselines, ridge regression achieves the highest temporal R^2 of 0.6490, closely followed by linear regression at 0.6477. The two tree-based models exhibit a pronounced decline: random forest temporal R^2 drops to 0.5103 and GBRT to 0.4865, despite their cross-validation RMSE being comparable to the linear models.

The convolutional models maintain substantially higher temporal R^2 values, ranging from 0.7597 (Dual-Branch) to 0.7674 (Middle Fusion). Middle Fusion achieves the highest temporal R^2 of 0.7674, exceeding the best tabular result (Ridge, 0.6490) by an absolute gap of 0.118 in R^2 and reducing temporal RMSE by 18.6% (1.9517 vs. 2.3976 $\mathrm{gC\,m^{-2}\,d^{-1}}$). In this dataset, spatial structure at 10 m generalises more robustly across years than scalar spectral features.

4.3 Geographic holdout

Geographic transfer to the 18-site upper Midwest forest holdout is associated with a marked increase in RMSE and a decrease in R^2 across all models relative to the temporal setting. The best OOD R^2 across all seven models is 0.5866 (Early Fusion), compared to 0.7674 for the best temporal R^2 (Middle Fusion). The holdout region consists of dense deciduous and mixed forests in Wisconsin and Michigan, a spatially coherent cluster whose landscape structure differs from the geographically dispersed training pool (Figure 2). The holdout also exhibits a substantially higher mean GPP (5.67 $\mathrm{gC\,m^{-2}\,d^{-1}}$) than the training pool (3.06 $\mathrm{gC\,m^{-2}\,d^{-1}}$), adding a target-distribution shift to the structural difference from the training pool.

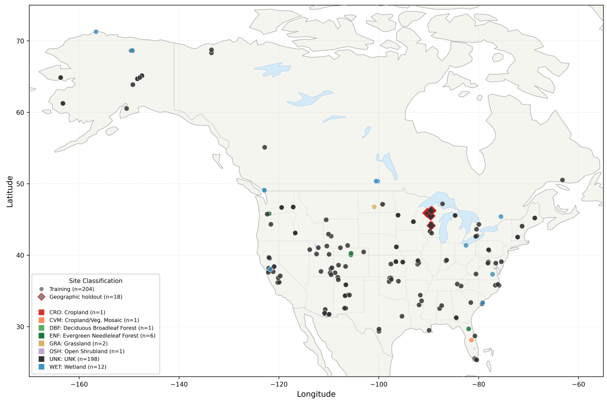

Figure 2. Spatial distribution of the 222 AmeriFlux sites used in this study. The 18-site upper Midwest forest holdout cluster (red diamonds) is geographically separated from the remaining 204 training-pool sites (black circles).

Within this setting, the convolutional models still outperform all tabular baselines. Early Fusion achieves the highest OOD R^2 of 0.5866 with an RMSE of 3.1509 $\mathrm{gC\,m^{-2}\,d^{-1}}$, while the best tabular model (Linear, $R^2 = 0.5371$, $\mathrm{RMSE} = 3.3346$) falls 0.050 below in R^2. The gap between the model families is narrower than under temporal generalisation (0.050 vs. 0.118 in R^2); the spatial advantage diminishes when the target region differs structurally from the training distribution. Among the tabular baselines, the linear models again outperform the tree ensembles, with random forest and GBRT R^2 values of 0.4840 and 0.4681 respectively, matching the temporal ranking and confirming limited robustness under distributional shift.

The ordering within the convolutional family shifts between the temporal and geographic settings. Middle Fusion, which achieves the highest temporal R^2 (0.7674), yields the lowest OOD R^2 among the three convolutional architectures (0.5480). Early Fusion, which ranks second in temporal R^2 (0.7647), yields the strongest OOD performance. These intra-family differences are small and are reported from a single seed.

Across both generalisation axes, the convolutional models that learn directly from pixel-level spatial patterns consistently outperform the tabular baselines that compress each patch to spectral aggregates. The advantage is most pronounced under temporal transfer, where the best convolutional model explains 77% of GPP variance compared to 65% for the best tabular baseline,

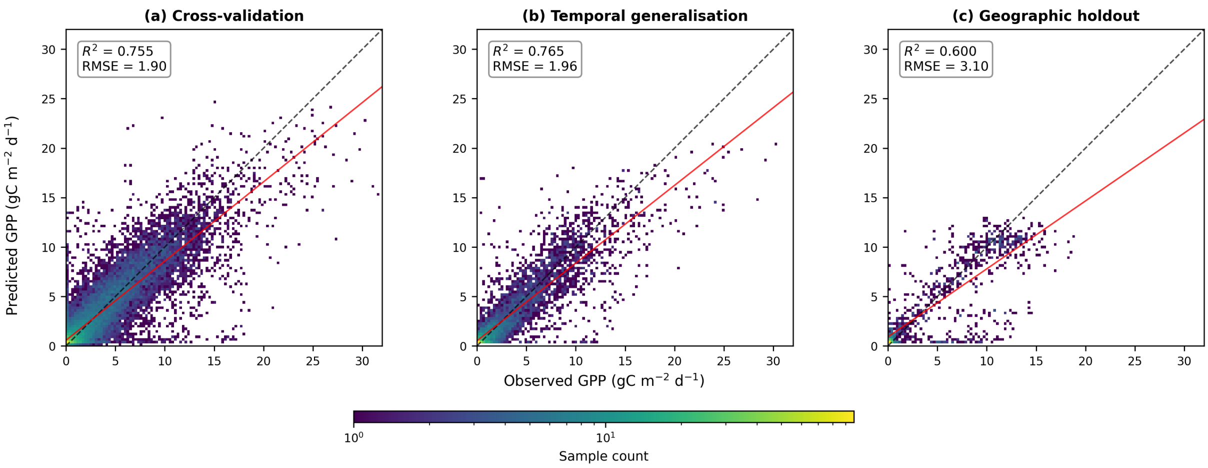

Figure 3. Predicted versus observed GPP for Early Fusion under (a) cross-validation, (b) temporal generalisation (train 2015–2021, test 2023–2025), and (c) geographic holdout. Colour encodes sample density on a logarithmic scale; the dashed line marks 1:1 agreement and the solid line shows a linear fit. For panel (c), predictions are averaged across the five geographic-holdout folds before plotting, whereas Table 2 reports the mean and standard deviation of the per-fold metrics.

and persists under geographic transfer, where all three convolutional architectures exceed all four tabular models in R^2. For Early Fusion, prediction density remains concentrated near the 1:1 line under cross-validation and temporal transfer, but becomes more dispersed under geographic holdout, particularly at high GPP values (Figure 3).

5. Discussion

Temporal transfer results indicate that two-dimensional spatial structure provides a more temporally stable basis for GPP estimation than spectral aggregates. Convolutional models trained on 2015–2021 data explain 77% of GPP variance in 2023–2025, compared with 65% for the best tabular baseline (Ridge). The landscape properties that convolutional models learn at 10 m resolution (canopy architecture, land-cover boundaries, within-stand heterogeneity) change slowly relative to the interannual variability in atmospheric and phenological conditions that drives spectral reflectance. The tabular baselines, which compress each patch to $\sim$38 scalar features, are directly exposed to these year-to-year distributional shifts. The fusion of optical and radar inputs also combines complementary information: Sentinel-2 reflectance captures biochemical vegetation traits, while Sentinel-1 backscatter encodes physical structural properties such as surface roughness and canopy moisture.

This spatial advantage, however, contracts under geographic transfer. Although all three convolutional architectures still outperform all four tabular baselines in the upper Midwest forest holdout (best OOD $R^2 = 0.59$ vs. 0.54), the margin is narrower than under the temporal setting ($\Delta R^2 = 0.050$ vs. 0.118). The spatial patterns that stabilise temporal predictions themselves differ between the holdout and the geographically dispersed training pool, so the spatial anchor provides less leverage when the target landscape structure is novel. The holdout's closed-canopy deciduous stands exhibit a pronounced phenological cycle (dense leaf-on canopy in summer, near-complete leaf-off in winter) that contrasts with the more heterogeneous mix of evergreen, grassland, and cropland sites in the training data.

Model capacity alone does not explain the convolutional advantage. Under temporal transfer, ridge regression, with $\sim$40 learnable coefficients, outperforms random forest (300 trees, unlimited depth) and GBRT (500 trees) by 0.14–0.16 R^2, showing that a larger parameter budget does not improve transfer by itself. The tree ensembles partition the feature space into axis-aligned decision boundaries calibrated to the training-period reflectance distributions, which fail to transfer when those distributions shift. The near-proportionality between vegetation indices and light absorption (Running et al., 2004, Pei et al., 2022) renders the spectral–GPP relationship approximately linear, explaining why the simplest tabular model transfers most robustly. The convolutional models also benefit from geometric augmentation and modal dropout, techniques made possible by retaining the two-dimensional image layout. Retaining image patches therefore provides both spatial context and augmentation-based regularisation.

Despite these advantages, the narrowing performance under geographic transfer exposes a limitation of the current input design. All models rely exclusively on Sentinel-1 and Sentinel-2 imagery with temporal and positional encoding; no meteorological variables, phenology indices, or ancillary satellite products are incorporated. GPP is driven not only by landscape structure but also by incoming radiation, temperature, vapour-pressure deficit, and soil moisture, none of which are directly represented in the input tensor. Integrating ancillary environmental fields or derived phenological metrics may help reduce this gap by supplying context not present in Sentinel imagery alone.

6. Conclusions

A systematic evaluation of 8-day GPP estimation from 10 m Sentinel-1 and Sentinel-2 imagery was presented, comprising 23,528 quality-controlled samples from 222 AmeriFlux sites over 2015–2025, with an 18-site upper Midwest forest holdout for out-of-distribution evaluation. Four tabular baselines and three convolutional fusion architectures were compared under cross-validation, temporal generalisation, and geographic transfer.

Two-dimensional spatial structure at 10 m resolution, drawing on complementary optical and radar inputs, yields measurable gains over spectral summary statistics, and the advantage is strongest under temporal transfer. The best convolutional model trained on 2015–2021 data explains 77% of GPP variance in 2023–2025 ($R^2 = 0.77$, RMSE = $1.95\,\mathrm{gC\,m^{-2}\,d^{-1}}$), compared with 65% for ridge regression ($R^2 = 0.65$, RMSE = $2.40\,\mathrm{gC\,m^{-2}\,d^{-1}}$), an 18.6% reduction in RMSE. Among the tabular baselines, linear models transfer more robustly than tree ensembles. Under geographic transfer to the upper Midwest forest holdout, the convolutional advantage persists but is smaller ($R^2 = 0.59$ vs. 0.54), as landscape-specific spatial patterns differ between training and target domains.

Evaluating temporal and geographic transfer alongside cross-validation on a common dataset reveals where spatial information adds value and where domain-dependent gaps remain. The separation between model families emerges most clearly under temporal transfer, not cross-validation, underscoring the need for multi-axis evaluation when assessing the transferability of satellite-derived GPP models beyond their training distribution. Establishing such robust spatial representations from high-resolution multi-sensor imagery is a step toward reliable carbon monitoring across heterogeneous landscapes.

Acknowledgements

The financial support from the Canadian Space Agency (CSA) [24AO3YOR35] and the Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2021-03624] is gratefully acknowledged. AmeriFlux site principal investigators, site teams, and their funding agencies are acknowledged for producing and sharing eddy-covariance flux observations. AmeriFlux data were made available through the data portal (`https://ameriflux.lbl.gov/`) and processing maintained by the AmeriFlux Management Project, supported by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research, under contract number DE-AC02-05CH11231.

Data Availability

AmeriFlux eddy-covariance data are available from the AmeriFlux data portal (`https://ameriflux.lbl.gov/`) under site-specific data-use agreements. Sentinel-2 Level-2A and Sentinel-1 RTC imagery were accessed through the Microsoft Planetary Computer STAC API and are freely accessible through the Copernicus Open Access Hub.

References

Alaska Satellite Facility, n.d. Sentinel-1 RTC product guide. `https://hyp3-docs.asf.alaska.edu/guides/rtc_product_guide/`. Accessed: 2026-03-28.

AmeriFlux, 2021. Flux data products. `https://ameriflux.lbl.gov/data/flux-data-products/`. Accessed: 2026-03-28.

Baldocchi, D., Novick, K., Keenan, T., Torn, M., 2024. AmeriFlux: Its Impact on Our Understanding of the ‘Breathing of the Biosphere’, after 25 Years. *Agricultural and Forest Meteorology*, 348, 109929. https://doi.org/10.1016/j.agrformet.2024.109929.

Bi, W., He, W., Zhou, Y., Ju, W., Liu, Y., Liu, Y., Zhang, X., Wei, X., Cheng, N., 2022. A Global 0.05-Degree Dataset for Gross Primary Production of Sunlit and Shaded Vegetation Canopies from 1992 to 2020. *Scientific Data*, 9(1), 213. https://doi.org/10.1038/s41597-022-01309-2.

Cai, Z., Junttila, S., Holst, J., Jin, H., Ardö, J., Ibrom, A., Peichl, M., Mölder, M., Jönsson, P., Rinne, J., Karamihalaki, M., Eklundh, L., 2021. Modelling Daily Gross Primary Productivity with Sentinel-2 Data in the Nordic Region–Comparison with Data from MODIS. *Remote Sensing*, 13(3), 469. https://doi.org/10.3390/rs13030469.

Chu, H., Luo, X., Ouyang, Z., Chan, W. S., Dengel, S., Biraud, S. C., Torn, M. S., Metzger, S., Kumar, J., Arain, M. A. et al., 2021. Representativeness of Eddy-Covariance Flux Footprints for Areas Surrounding AmeriFlux Sites. *Agricultural and Forest Meteorology*, 301–302, 108350. https://doi.org/10.1016/j.agrformet.2021.108350.

Copernicus Sentinel Wiki, n.d. Sentinel-2 processing. `https://sentiwiki.copernicus.eu/web/s2-processing`. Accessed: 2026-03-28.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P. et al., 2012. Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, 120, 25–36. https://doi.org/10.1016/j.rse.2011.11.026.

FLUXNET, n.d. FLUXNET2015 variables quick start guide. `https://fluxnet.org/data/fluxnet2015-dataset/variables-quick-start-guide/`. Accessed: 2026-03-23.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Joshi, N., Baumann, M., Ehammer, A., Fensholt, R., Grogan, K., Hostert, P., Jepsen, M., Kuemmerle, T., Meyfroidt, P., Mitchard, E., Reiche, J., Ryan, C., Waske, B., 2016. A Review of the Application of Optical and Radar Remote Sensing Data Fusion to Land Use Mapping and Monitoring. *Remote Sensing*, 8(1), 70. https://doi.org/10.3390/rs8010070.

Jung, M., Schwalm, C., Migliavacca, M., Walber, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N. et al., 2020. Scaling Carbon Fluxes from Eddy Covariance Sites to Globe: Synthesis and Evaluation of the FLUXCOM Approach. *Biogeosciences*, 17(5), 1343–1365. https://doi.org/10.5194/bg-17-1343-2020.

Junttila, S., Ardö, J., Cai, Z., Jin, H., Kljun, N., Klemedtsson, L., Krasnova, A., Lange, H., Lindroth, A., Mölder, M., Noe, S. M., Tagesson, T., Vestin, P., Weslien, P., Eklundh, L., 2023. Estimating Local-Scale Forest GPP in Northern Europe Using Sentinel-2: Model Comparisons with LUE, APAR, the Plant Phenology Index, and a Light Response Function. *Science of Remote Sensing*, 7, 100075. https://doi.org/10.1016/j.srs.2022.100075.

Kong, J., Ryu, Y., Liu, J., Dechant, B., Rey-Sanchez, C., Shortt, R., Szutu, D., Verfaillie, J., Houborg, R., Baldocchi, D. D., 2022. Matching High Resolution Satellite Data and Flux Tower Footprints Improves Their Agreement in Photosynthesis Estimates. *Agricultural and Forest Meteorology*, 316, 108878. https://doi.org/10.1016/j.agrformet.2022.108878.

LAADS DAAC, 2015. MODIS/Terra Gross Primary Productivity 8-Day L4 Global 500m SIN Grid (MOD17A2H). `https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MOD17A2H/`. Accessed: 2026-03-23.

Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., Yosinski, J., 2018. An intriguing failing of convolutional neural networks and the CoordConv solution. *Advances in Neural Information Processing Systems*, 31, 9605–9616.

Pabon-Moreno, D. E., Migliavacca, M., Reichstein, M., Mahecha, M. D., 2022. On the Potential of Sentinel-2 for Estimating Gross Primary Production. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–12. https://doi.org/10.1109/TGRS.2022.3152272.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W. et al., 2020. The FLUXNET2015 Dataset and the ONEFlux Processing Pipeline for Eddy Covariance Data. *Scientific Data*, 7(1), 225. https://doi.org/10.1038/s41597-020-0534-3.

Pei, Y., Dong, J., Zhang, Y., Yuan, W., Doughty, R., Yang, J., Zhou, D., Zhang, L., Xiao, X., 2022. Evolution of Light Use Efficiency Models: Improvement, Uncertainties, and Implications. *Agricultural and Forest Meteorology*, 317, 108905. https://doi.org/10.1016/j.agrformet.2022.108905.

Running, S. W., Nemani, R. R., Heinsch, F. A., Zhao, M., Reeves, M., Hashimoto, H., 2004. A Continuous Satellite-Derived Measure of Global Terrestrial Primary Production. *BioScience*, 54(6), 547–560. https://doi.org/10.1641/0006-3568(2004)054[0547:ACSMOG]2.0.CO;2.

Spinosa, A., Fuentes-Monjaraz, M. A., El Serafy, G., 2023. Assessing the Use of Sentinel-2 Data for Spatio-Temporal Upscaling of Flux Tower Gross Primary Productivity Measurements. *Remote Sensing*, 15(3), 562. https://doi.org/10.3390/rs15030562.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M. et al., 2012. GMES Sentinel-1 Mission. *Remote Sensing of Environment*, 120, 9–24. https://doi.org/10.1016/j.rse.2011.05.028.

Vira, J., Vekuri, H., Nevalainen, O., Korkiakoski, M., Mattila, T., Aaltonen, H., Koskinen, M., Lohila, A., Pihlatie, M., Liski, J., 2025. Improving Agricultural Carbon Monitoring with Sentinel-2 and Eddy-Covariance-Based Plant Productivity Estimates. *Carbon Management*, 16(1), 2568042. https://doi.org/10.1080/17583004.2025.2568042.

Wolanin, A., Camps-Valls, G., Gómez-Chova, L., Mateo-García, G., van der Tol, C., Zhang, Y., Guanter, L., 2019. Estimating Crop Primary Productivity with Sentinel-2 and Landsat 8 Using Machine Learning Methods Trained with Radiative Transfer Simulations. *Remote Sensing of Environment*, 225, 441–457. https://doi.org/10.1016/j.rse.2019.03.002.

Zhu, W., Xie, Z., Zhao, C., Zheng, Z., Qiao, K., Peng, D., Fu, Y. H., 2024. Remote Sensing of Terrestrial Gross Primary Productivity: A Review of Advances in Theoretical Foundation, Key Parameters and Methods. *GIScience & Remote Sensing*, 61(1), 2318846. https://doi.org/10.1080/15481603.2024.2318846.