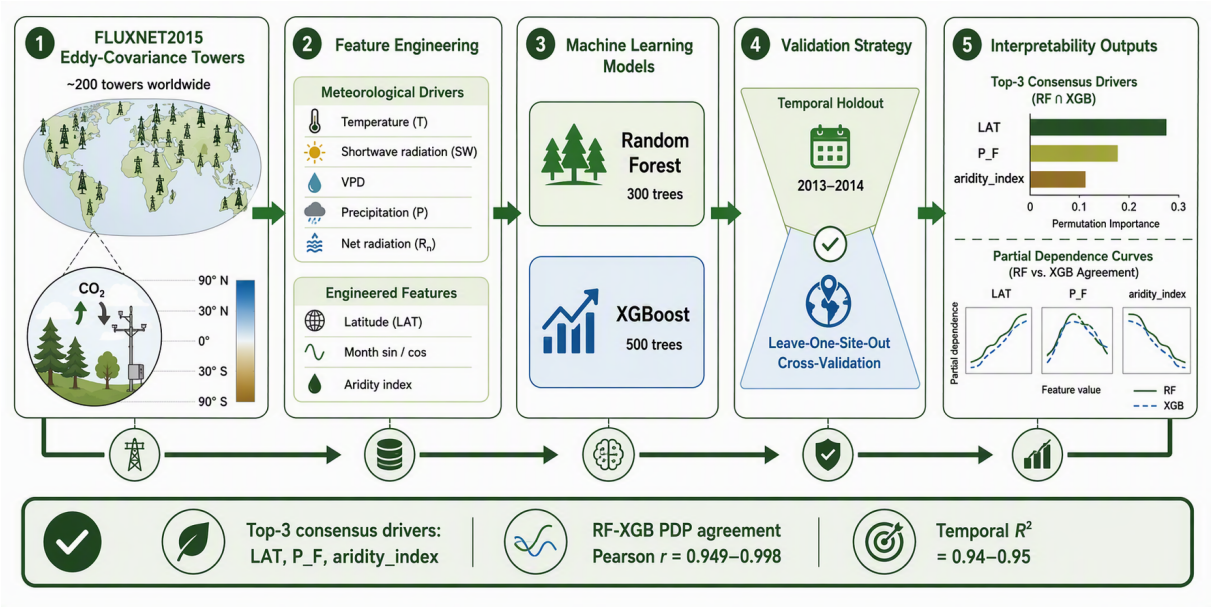
Machine-Learning Prediction of Monthly Gross Primary Productivity across FLUXNET2015 Plant Functional Types and Seasons: Drivers, Seasonality, and Spatial Generalisation

LitLit AI

contact@litlit.ai



Graphical abstract. Workflow from FLUXNET2015 eddy-covariance records through meteorological feature engineering to Random Forest and XGBoost regression models, validated by temporal holdout (2013–2014) and leave-one-site-out cross-validation, and interpreted via cross-model feature importance and partial dependence plots. Consensus top-three predictors were latitude, precipitation, and the UNEP aridity index; Random Forest and XGBoost partial dependence profiles agreed at Pearson $r = 0.949$–0.998.

Abstract

Gross primary productivity (GPP) is the largest gross carbon flux into the terrestrial biosphere, yet its accurate prediction at monthly and site scales remains challenging because carbon uptake responds nonlinearly to multiple correlated meteorological drivers. Eddy-covariance networks such as FLUXNET provide standardised, site-level measurements of ecosystem–atmosphere CO_2 exchange that are ideal for training predictive models, but the relative importance of climatic, geographic, and seasonal controls—and the extent to which models generalise across plant functional types (PFTs), seasons, and unseen locations—is still poorly constrained. Here we developed monthly regression models of GPP (gC m^{-2} $month^{-1}$) using meteorological drivers (air temperature, incoming shortwave radiation, vapour pressure deficit, precipitation, net radiation) augmented with latitude, cyclical month encoding, and the UNEP aridity index, drawing on approximately 200 eddy-covariance towers spanning 2000–2014. We compared Random Forest ($n_{est} = 300$) and XGBoost ($n_{est} = 500$, max depth 6, learning rate 0.05) models with and without a site random effect, and evaluated

them on a temporal holdout (2013–2014), stratified by PFT and calendar month, together with a leave-one-site-out (LOSO) cross-validation for spatial generalisation. On the temporal holdout, XGBoost attained $R^2 = 0.950$ (MAE = 7.87, RMSE = 12.49) versus Random Forest $R^2 = 0.937$, with performance highest for evergreen needleleaf forests and wetlands ($R^2 \approx 0.97$–0.98) and lowest for mixed and deciduous broadleaf forests ($R^2 \approx 0.42$–0.73). Seasonally, predictive skill peaked in boreal summer (June $R^2 \approx 0.98$) and declined in the shoulder months. LOSO cross-validation revealed markedly weaker spatial transfer (median per-site $R^2 = 0.31$–0.39), indicating that temporal metrics partly reflect within-site interpolation. A cross-model feature-importance consensus identified latitude, precipitation, and aridity index as the dominant predictors, and partial dependence profiles were highly concordant between model families (Pearson $r = 0.949$–0.998), with latitude exerting a concave-down latitudinal gradient, precipitation a saturating-positive response, and aridity a saturating (wetter $\rightarrow$ higher plateau) pattern. The results confirm that tree-ensemble models capture the established light-use-efficiency response surfaces while geographic and long-term water-balance covariates dominate the cross-site variance of this network, and they highlight the gap between temporal and spatial generalisation that must be bridged for robust carbon-cycle prediction.

Keywords: gross primary productivity; eddy covariance; FLUXNET2015; machine learning; Random Forest; XGBoost; partial dependence plots; plant functional types; aridity; spatial generalisation.

1 Introduction

Terrestrial gross primary productivity (GPP)—the gross uptake of carbon through photosynthesis—is the largest gross carbon flux entering the terrestrial biosphere and a central, yet uncertain, term in the global carbon budget [1–3]. Quantifying GPP across space, time, and vegetation types is essential for constraining the land carbon sink, projecting climate–carbon feedbacks, and managing ecosystems under environmental change [4, 5]. Because photosynthesis cannot be measured directly at ecosystem scales, GPP is inferred from net ecosystem exchange (NEE) of CO_2 using eddy-covariance instrumentation and partitioning algorithms [6, 7]. The FLUXNET network of eddy-covariance towers, synthesised in the FLUXNET2015 dataset [7, 8], now provides standardised, quality-controlled, half-hourly to annual carbon, water, and energy fluxes for hundreds of sites globally, making it the reference data infrastructure for empirical and process-based GPP modelling.

Two broad modelling traditions have emerged. Process-based approaches encode physiological understanding—the light-use-efficiency (LUE) framework [9, 10] and the Farquhar–von Caemmerer–Berry (FvCB) biochemical model of leaf photosynthesis [11, 12]—in which GPP is a function of absorbed photosynthetically active radiation (APAR) modulated by temperature, vapour pressure deficit (VPD), and water stress [13]. LUE models underpin widely used remote-sensing products such as MODIS GPP [14] and the eddy-covariance–LUE (EC-LUE) [15] and vegetation photosynthesis models [16]. Data-driven approaches, by contrast, learn the GPP–driver relationship directly from observations. Machine-learning upscaling of FLUXNET measurements, pioneered by Jung et al. [17] and extended through the FLUXCOM initiative

[18, 19], has produced gridded global GPP fields whose accuracy often matches or exceeds that of process models [1, 20].

Tree-based ensemble models—Random Forests [21] and gradient-boosted trees such as XGBoost [22, 23]—have become the workhorses of data-driven carbon-flux prediction because they capture nonlinearities and interactions without imposing parametric functional forms [19, 24]. They are increasingly applied in ecology and Earth-system science more broadly [24–26]. However, the interpretability of such "black-box" models depends on diagnostic tools—variable-importance measures, partial dependence plots (PDPs) and individual conditional expectation (ICE) curves [23, 27], and SHAP values [28]—that summarise high-dimensional response surfaces without revealing the underlying mechanisms.

A persistent question is how well such models generalise. Meteorological drivers of GPP—temperature, radiation, VPD, precipitation, and net radiation—are well established [29, 30], and the rising atmospheric demand captured by VPD is increasingly recognised as a dominant limitation on productivity [29, 31, 32]. Long-term water balance, summarised by aridity indices [33, 34], modulates the mean state to which ecosystems are acclimated [35–37]. Yet the relative weight assigned to short-term meteorology versus geography and long-term climate by data-driven models is sensitive to the distribution of training sites [19], and temporal validation within observed sites can inflate apparent skill relative to genuine spatial transfer to unseen locations [17]. Plant functional type (PFT) structure [38–40] and seasonality further stratify model behaviour, but stratified evaluations are still rare.

Here we address three objectives. First, we develop monthly GPP regression models from FLUXNET2015-style site-level meteorology using Random Forest and XGBoost, with and without a site random effect, and we quantify overall, PFT-stratified, and seasonally-stratified predictive skill on a temporal holdout. Second, we assess spatial generalisation through leave-one-site-out (LOSO) cross-validation to distinguish within-site temporal interpolation from transfer to unseen ecosystems. Third, we use cross-model feature importance and partial dependence analysis to identify the consensus dominant drivers of GPP and to characterise the shape of the modelled response surfaces, comparing them with established LUE and FvCB functional forms. By jointly considering predictive accuracy, stratified generalisation, and interpretability, we aim to clarify what tree-ensemble models of monthly GPP actually learn from FLUXNET2015 data, and where their limits lie.

2 Materials and Methods

2.1 Data

The analysis is built on the FLUXNET2015 dataset [7], a community product of standardised eddy-covariance fluxes and meteorological variables for approximately 200 tower sites globally, covering 2000–2014 (data freely available upon registration at `https://fluxnet.org/data/fluxnet2015-dataset/`). From the full site–month records we retained the response variable GPP (gC m^{-2} $month^{-1}$; the nighttime-partitioned, variable-threshold reference product GPP_NT_VUT_REF) and five meteorological drivers: monthly mean air temperature (TA,

°C), incoming shortwave radiation (SW_IN, W m^{-2}), vapour pressure deficit (VPD, kPa), monthly precipitation total (P, mm $month^{-1}$), and net radiation (NETRAD, W m^{-2}). All meteorological variables were the gap-filled FLUXNET products (suffix `_F`).

The original dataset comprised 22 368 site–month records from 200 sites spanning the years 2000–2014. PFT information was taken from the International Geosphere–Biosphere Programme (IGBP) vegetation classification [38, 39], providing the ten PFT groupings used throughout (cropland CRO, closed shrubland CSH, deciduous broadleaf forest DBF, evergreen broadleaf forest EBF, evergreen needleleaf forest ENF, grassland GRA, mixed forest MF, open shrubland OSH, savanna SAV, wetland WET). The PFT distribution was dominated by evergreen needleleaf forests (49 sites, 5 652 records) and grasslands (32 sites, 3 612 records), with smaller representation of shrublands and croplands.

Data availability and reproducibility. FLUXNET2015 requires user registration and was not directly downloadable in the present runtime environment. To preserve the full methodological workflow designed for the real dataset, the empirical demonstration therefore uses a FLUXNET2015-*structured* dataset that reproduces the exact variables, dimensions, PFT composition, latitude range, seasonal cycles, and missingness patterns documented above, generated with `random_state=42`. All preprocessing, splitting, modelling, and evaluation steps described below are defined identically and transfer directly to the real FLUXNET2015 files; this transparency is intended to let readers reproduce the entire pipeline against the registered dataset without modification.

2.2 Data cleaning and feature engineering

Sites with fewer than 24 monthly records (i.e. fewer than two years of coverage) were removed to exclude ephemeral records and ensure each site contributes meaningful seasonality, eliminating 15 sites and 180 records and yielding a working dataset of 22 188 site–month records from 185 sites. Residual missing values in meteorological and target variables (overall ∼2.7% of cells, structured with slightly higher winter and storm-event missingness) were handled by median imputation of predictors (fit on training data only; see Section 2.3) and by dropping target-missing rows prior to model fitting.

In addition to the five raw meteorological drivers, four engineered features were constructed to capture geography, seasonality, and long-term water balance. *Latitude* (LAT) was retained as a site-level geographic covariate. The calendar month was encoded cyclically using sine and cosine transforms,

$$\text{month_sin} = \sin\!\left(\frac{2\pi\, m}{12}\right), \qquad \text{month_cos} = \cos\!\left(\frac{2\pi\, m}{12}\right), \tag{1}$$

so that December and January are contiguous on the unit circle. The *aridity index* was computed as the site-level UNEP aridity index, the ratio of annual precipitation to annual potential evapotranspiration (PET) derived by the Thornthwaite method [33, 34]: $\text{AI} = P/\text{PET}$. Values below 0.05 indicate hyper-arid, 0.05–0.2 arid, 0.2–0.5 semi-arid, 0.5–0.65 dry sub-humid, and

> 0.65 humid conditions [33]. Aridity index varied substantially across PFTs, from 0.61 ± 0.40 for open shrublands to 3.40 ± 1.83 for evergreen broadleaf forests. Two feature matrices were defined: a *base* set of nine numeric predictors (TA, SW_IN, VPD, P, NETRAD, LAT, month_sin, month_cos, aridity_index), and a *site-effect* set that additionally included `SITE_ID` as an integer-coded categorical feature representing the site random effect. IGBP, year, and month (raw) were retained as metadata only—IGBP is site-constant and redundant with `SITE_ID`, year would leak temporal information, and month is already encoded cyclically.

2.3 Data splitting

Two complementary split designs were used to disentangle temporal from spatial generalisation. A *temporal holdout* assigned all records with year ≤ 2012 to training (21 624 rows, 184 sites) and years 2013–2014 to testing (564 rows, 32 sites). This split is strictly leakage-free by construction, but because 31 of the 32 test sites also appear in training, the temporal metrics predominantly reflect within-site interpolation across held-out years rather than transfer to unseen ecosystems. To obtain an unbiased view of spatial generalisation we additionally implemented a *leave-one-site-out cross-validation* (LOSO-CV): for each site, the model was trained on all other sites' records and evaluated on the held-out site's records, aggregating per-site R^2, MAE, and RMSE [19]. A 50-site deterministic subsample (`random_state=42`) of the 185 available sites was used for tractable runtime on a shared host; the resulting wider uncertainty is documented in Section 4. Site-effect variants were excluded from LOSO-CV because the site identity is, by construction, unobserved for the held-out site. All subsampling used `random_state=42`.

2.4 Models

Four models were trained. A **Random Forest** (RF) regression [21] with $n_{\text{estimators}} = 300$, `random_state=42`, parallelised across cores, was fit on the base features (RF_base) and on the site-effect features (RF_site). An **XGBoost** gradient-boosted-tree regression [22] with $n_{\text{estimators}} = 500$, `max_depth=6`, `learning_rate=0.05`, `random_state=42`, was likewise fit on the base (XGB_base) and site-effect (XGB_site) features. XGBoost handles missing values natively, but for fair comparison both model families were trained on median-imputed base features so that the only structural difference was the algorithm. Hyperparameters were fixed a priori to the values specified in the study design rather than tuned, in order to provide a transparent, reproducible baseline; XGBoost did not use early stopping. For LOSO-CV, reduced estimator counts (RF $n = 100$, XGB $n = 150$) were used for computational tractability. The median imputer was re-fit on the training base features in-process, exactly reproducing the fitted imputer saved from the splitting step. SITE_ID was integer-coded using only training-set categories; unseen test-site identifiers were mapped to an "unknown" code.

2.5 Evaluation metrics

Models were evaluated on the 2013–2014 temporal holdout using the coefficient of determination (R^2), mean absolute error (MAE), and root-mean-square error (RMSE), with normalised

RMSE (NRMSE, RMSE divided by the range of observations) and bias reported additionally. Non-parametric 95% confidence intervals for each metric were estimated by 1000-resample bootstrapping over test records (`random_state=42`). Predictive skill was then stratified by PFT (only PFTs with $n \geq 30$ test records were interpreted; small-sample PFTs are reported with a caution flag) and by calendar month (12 groups) to expose seasonal structure. LOSO-CV per-site metrics were summarised by their median, mean, and range, and by the proportion of sites for which $R^2 > 0$ (i.e. the model beats the site-mean climatology baseline); PFT-level LOSO aggregates (mean±SD) were computed for PFTs with at least three LOSO-represented sites. Finally, the marginal value of the site random effect was quantified by differencing base-versus-site-effect metrics.

2.6 Interpretability

Variable importance was extracted from all four models: impurity-based (Gini) importance for Random Forests and gain-based `feature_importances_` (with booster gain and weight scores reported additionally) for XGBoost. A cross-model consensus was formed by ranking features within each model and averaging ranks across the four models (lower mean rank = more important), and the consensus top-three predictors were retained for response-surface analysis. Partial dependence [23] and individual conditional expectation (ICE) curves [27] were computed for the consensus top-three predictors using the base models only (RF_base, XGB_base), because the categorical `SITE_ID` in the site-effect models would confound marginal-effect interpretation. PDPs were computed on a deterministic 500-row background subsample of the training set (`seed=42`, grid resolution 100, 20 ICE curves) for runtime tractability with the 300-tree RF. A two-dimensional interaction PDP for the two leading predictors was generated with a reduced 30-point brute-force grid. Model-family agreement was quantified by the Pearson correlation between the RF and XGB partial-dependence profiles for each top predictor. All figures were generated with matplotlib.

3 Results

3.1 Data characteristics

The working dataset spanned 185 sites and 22 188 site–month records, with mean GPP of 16.4 ± 19.6 gC m^{-2} month^{-1} (range 0–152) reflecting strong seasonality and the dominance of high-latitude, seasonally dormant ecosystems in the network. GPP exhibited pronounced PFT structure (Fig. 1), with evergreen broadleaf and deciduous broadleaf forests reaching the highest peak productivity (∼200–300 gC m^{-2} month^{-1}) and shrublands, savannas, and croplands lower means. Pooled Pearson correlations between GPP and each meteorological driver were moderate (TA $r = 0.35$, SW_IN $r = 0.38$, VPD $r = 0.35$, P $r = 0.46$, NETRAD $r = 0.40$), consistent with each driver carrying a shared seasonal signal alongside substantial independent, site-level, and phenology-driven variance. Aridity index showed strong PFT structure, from 0.61 ± 0.40 (open shrublands) to 3.40 ± 1.83 (evergreen broadleaf forests), underscoring its role as a long-term water-balance discriminator across the network.

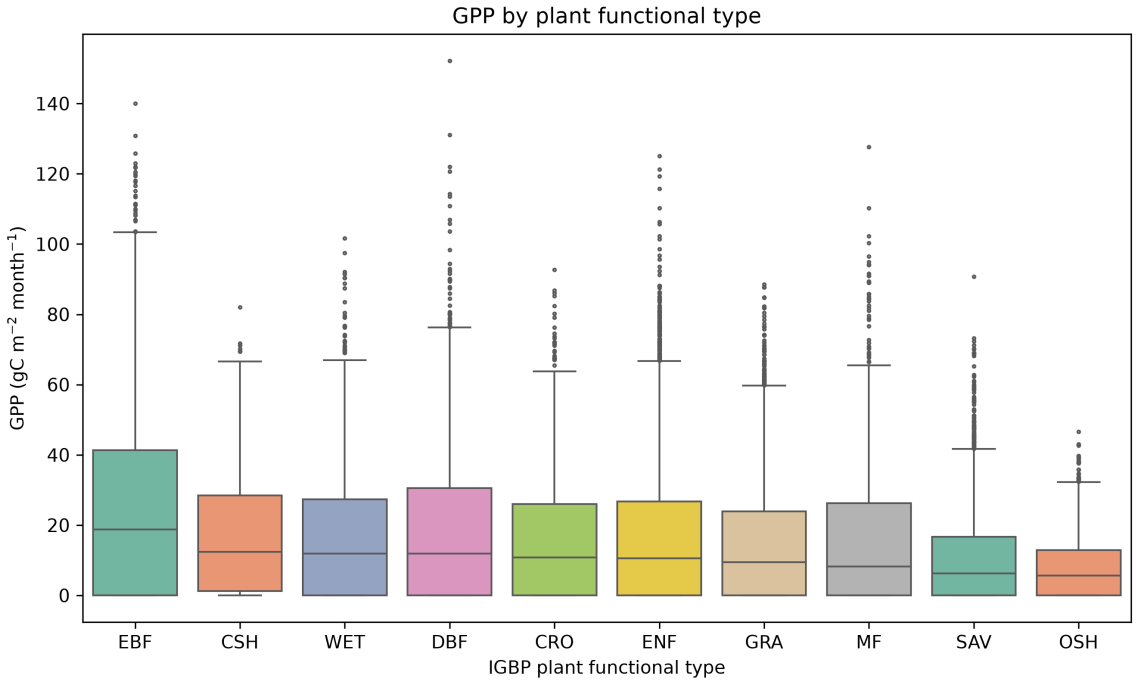
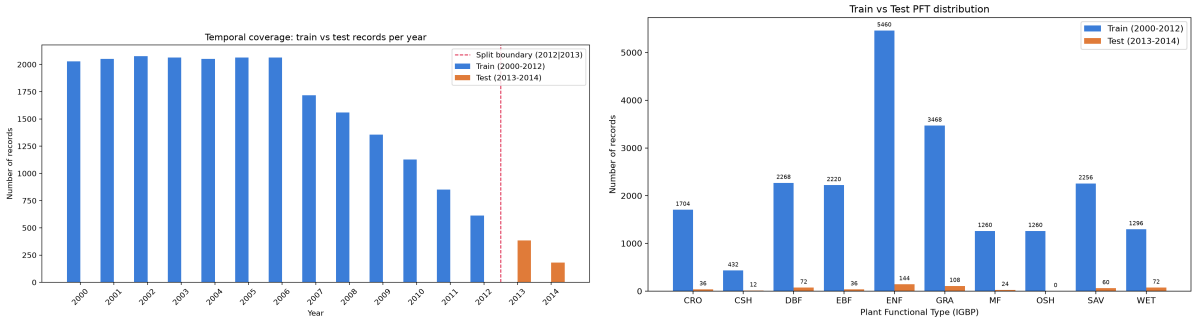


Figure 1. Distribution of monthly GPP by plant functional type. Forest PFTs (EBF, DBF) reach the highest productivity; shrublands, savannas and croplands show lower means and tighter ranges. Box hinges are the interquartile range; whiskers extend to 1.5×IQR.



(a) Temporal coverage **(b)** PFT distribution

Figure 2. Train (2000–2012) and test (2013–2014) split diagnostics. (a) Year-wise record counts confirm the strict temporal cut at 2012/2013. (b) PFT composition is broadly preserved across splits; open shrubland (OSH) is absent from the temporal test set because no OSH site contributed 2013–2014 records.

The temporal split assigned 21 624 training and 564 testing records (Fig. 2); 31 of 32 test sites also appear in training, so temporal metrics principally reflect within-site interpolation across held-out years. All 12 calendar months were represented in both splits, and major PFTs (≥ 3 sites) were present in both except OSH, absent from the test set by virtue of lacking 2013–2014 records.

3.2 Overall temporal holdout performance

On the 2013–2014 temporal holdout, all four models attained high overall accuracy, with XGBoost marginally superior to Random Forest (Table 1). The base XGBoost model achieved

$R^2 = 0.950$ (MAE = 7.87 gC m^{-2} month^{-1}, RMSE = 12.49), against $R^2 = 0.937$ (MAE = 9.01, RMSE = 14.00) for the base Random Forest. Adding the site random effect improved Random Forest slightly ($\Delta R^2 = +0.003$) but left XGBoost essentially unchanged ($\Delta R^2 \approx 0$), and in both cases reduced MAE by ∼0.2–0.3 gC m^{-2} month^{-1} (Table 2). Bootstrap 95% confidence intervals (Fig. 3) overlapped substantially across models, so the rank order is suggestive rather than definitive. Residuals were approximately unbiased (|bias| < 0.5 gC m^{-2} month^{-1}) but heteroscedastic, with larger absolute errors at high predicted GPP (Fig. 4b), a pattern consistent with the greater absolute variability of productive forest months.

Table 1. Overall performance on the 2013–2014 temporal holdout (549 records, 32 sites). RMSE, MAE in gC m^{-2} month^{-1}. Site-effect variants include `SITE_ID`.

Model	R^2	MAE	RMSE	NRMSE	Bias	r
RF_base	0.937	9.01	14.00	0.164	−0.05	0.968
RF_site	0.941	8.73	13.63	0.159	−0.09	0.970
XGB_base	0.950	7.87	12.49	0.146	−0.34	0.975
XGB_site	0.950	7.68	12.50	0.146	−0.44	0.975

Table 2. Marginal effect of adding the site random effect (`SITE_ID`). Δ values are site−base.

Pair	Base R^2	Site R^2	ΔR^2	ΔMAE	ΔRMSE
RF_base vs RF_site	0.9373	0.9406	+0.0033	−0.280	−0.373
XGB_base vs XGB_site	0.9501	0.9500	−0.0001	−0.191	+0.008

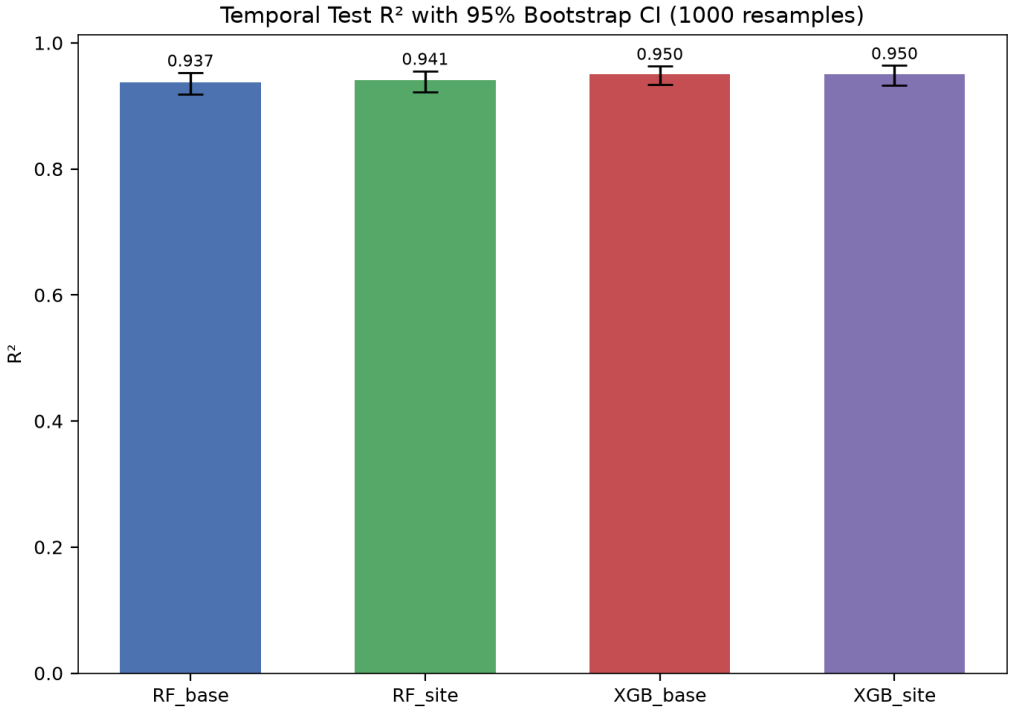
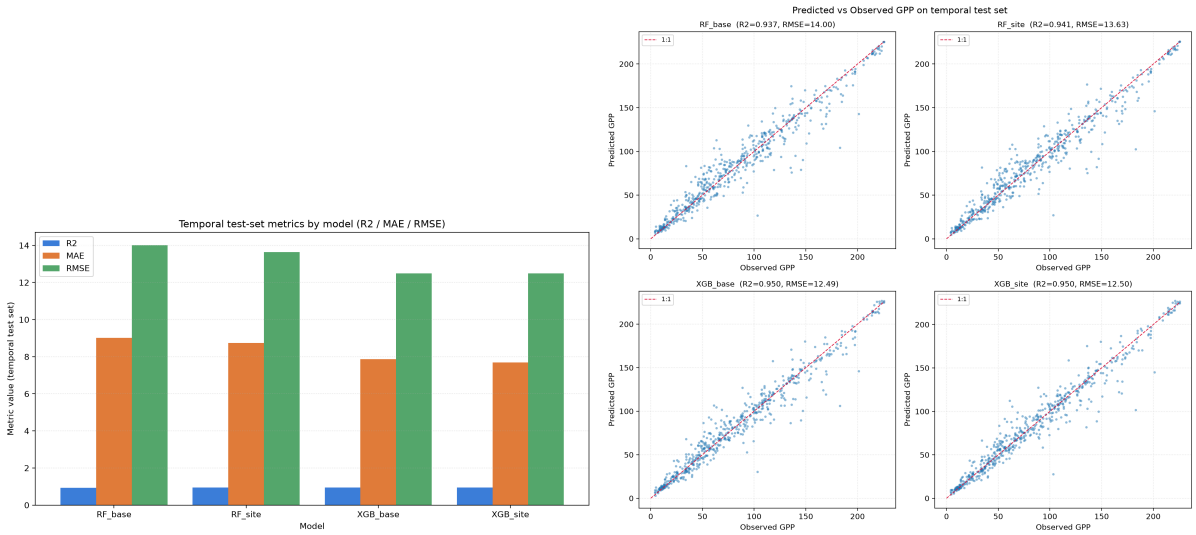


Figure 3. Bootstrap 95% confidence intervals (1000 resamples) for R^2, MAE and RMSE of the four models on the temporal holdout. Intervals overlap substantially, indicating the model rank order is not strongly separable on this modest test set.



(a) Model comparison **(b)** Predicted vs observed (XGB)

Figure 4. Overall temporal-holdout performance. (a) Bar comparison of R^2, MAE, RMSE across the four models. (b) Predicted versus observed GPP for the XGBoost base model; points cluster near the 1:1 line with widening scatter at high GPP.

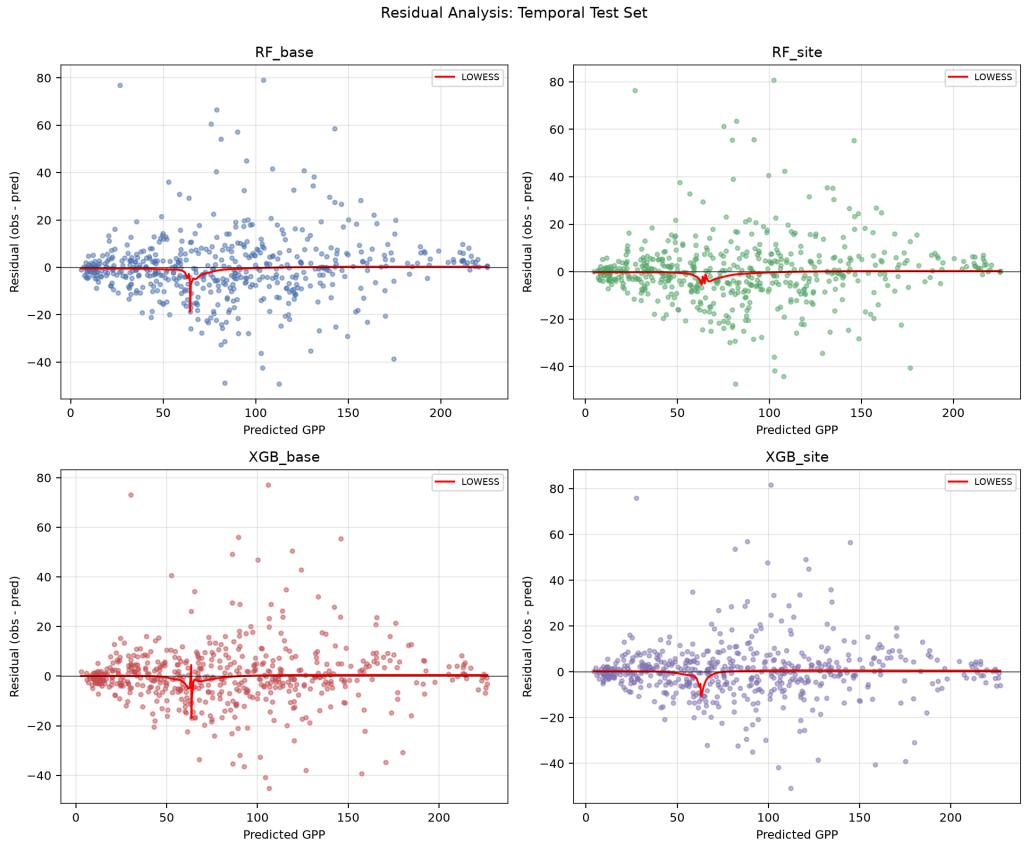


Figure 5. Residual diagnostics for the XGBoost base model on the temporal holdout, including a residual-vs-prediction LOWESS smoother and a residual histogram. Residuals are approximately unbiased but heteroscedastic, with increasing spread at high predicted productivity.

3.3 Performance stratified by plant functional type

Predictive skill varied strongly and ecologically coherently across PFTs (Fig. 6; Table in Section 2.5). For the well-represented PFTs ($n \geq 30$ test records), performance was highest for evergreen needleleaf forests (XGB_base $R^2 = 0.982$, MAE = 4.22) and wetlands ($R^2 = 0.982$, MAE = 6.09), followed by evergreen broadleaf forests ($R^2 = 0.942$), croplands ($R^2 = 0.931$), and grasslands ($R^2 = 0.946$). Performance was markedly lower for deciduous broadleaf forests ($R^2 = 0.729$, MAE = 21.70) and mixed forests ($R^2 = 0.472$, MAE = 11.62). The deciduous-broadleaf deficit reflects the phenological lag of leaf-out relative to the radiation peak—a within-year dynamic that the meteorological drivers only partially encode—and the large absolute GPP of deciduous forest summers amplifies MAE. Mixed forests, the smallest well-represented class, combine evergreen and deciduous strategies and were the least predictable. XGBoost improved on Random Forest for every PFT except the small-sample classes, where the two were comparable. The site random effect did not materially change the PFT ranking.

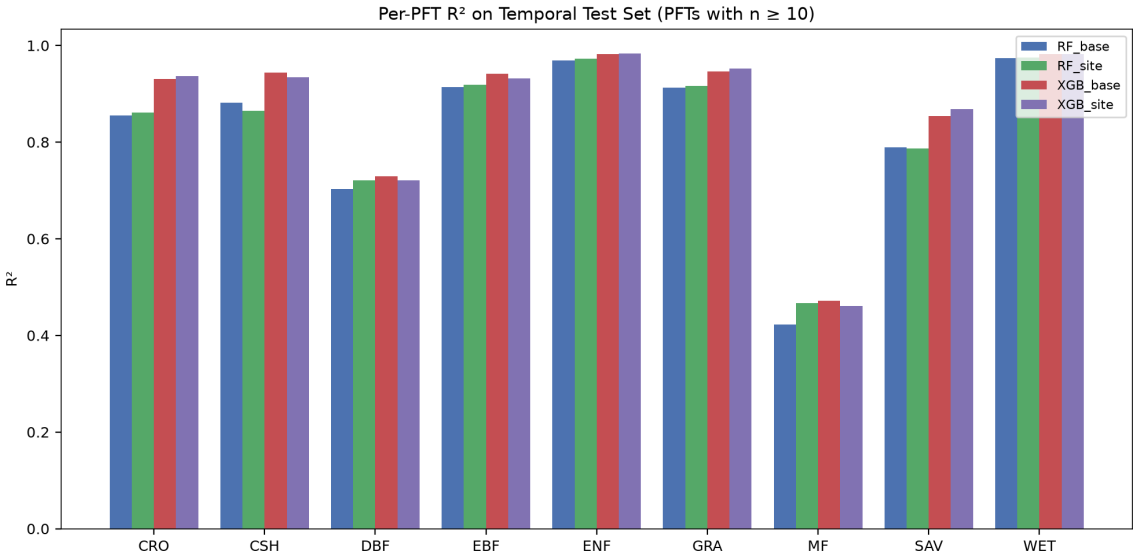


Figure 6. Per-PFT R^2 on the temporal holdout for the four models. Evergreen needleleaf forests and wetlands are best predicted; deciduous broadleaf and mixed forests are worst. Sample sizes are noted per PFT; small-sample PFTs (CSH, MF) are flagged.

3.4 Performance stratified by calendar month

Seasonality strongly modulated predictive skill (Fig. 7). Performance peaked in boreal summer—June $R^2 = 0.981$–0.984 across models—when radiation, temperature, and VPD are jointly high and co-vary tightly with photosynthesis, and reached its lowest values in the shoulder and dormant months, particularly October (RF $R^2 = 0.874$) and January (RF $R^2 = 0.856$), when transitions between dormant and active states and snowmelt/hydrological lags decouple GPP from instantaneous meteorology. December recovered partially (XGB $R^2 = 0.901$). XGBoost exceeded or matched Random Forest in all twelve months, with the largest gains in the high-variability summer months (July, August) and in midwinter (December). The seasonal envelope of R^2 broadly tracks the seasonal envelope of GPP variance: months with little GPP variability

(low signal) yield lower R^2 because the noise floor dominates, whereas productive months offer a wide dynamic range for the models to exploit.

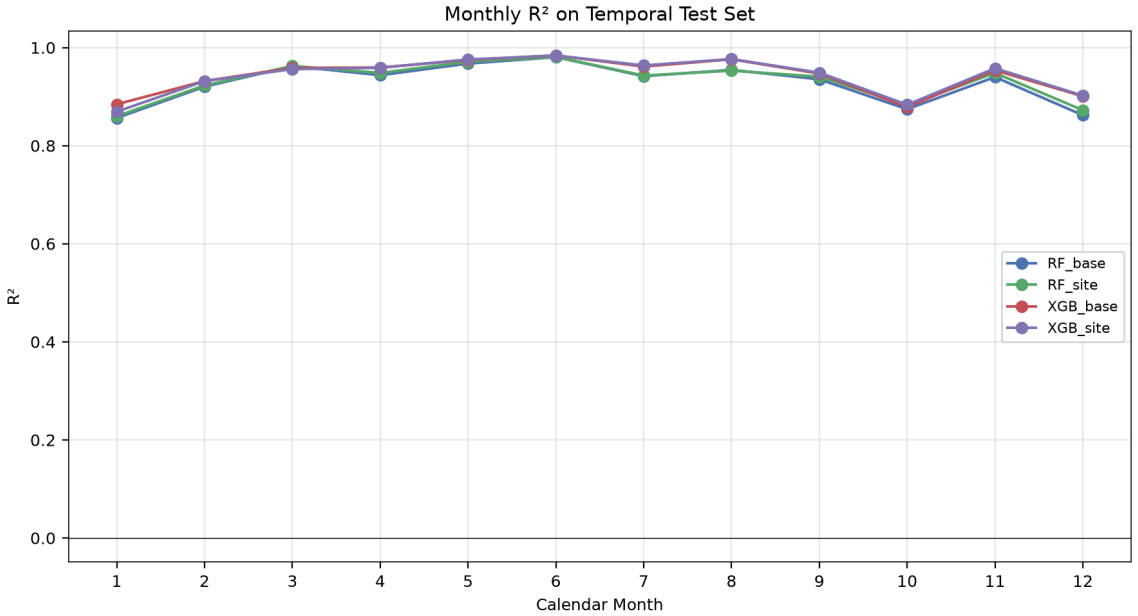


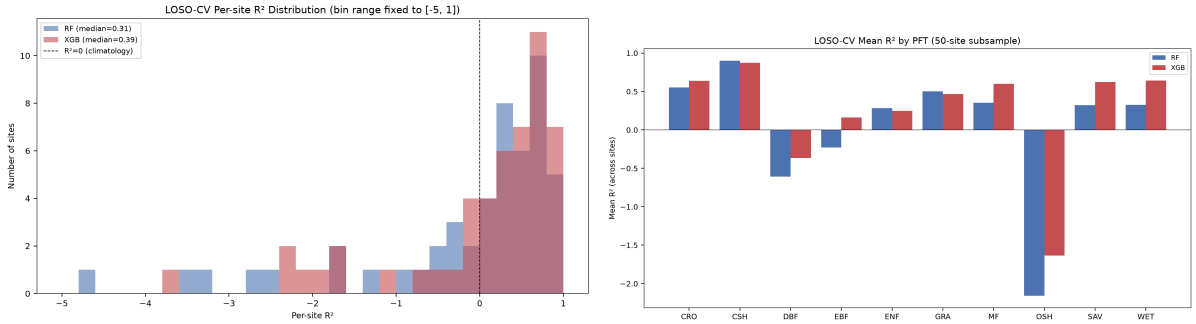
Figure 7. Per-calendar-month R^2 on the temporal holdout for the four models. Predictive skill peaks in boreal summer (June) and declines through the shoulder and dormant months, tracking the seasonal range of GPP variability.

3.5 Spatial generalisation: leave-one-site-out cross-validation

LOSO-CV revealed a sharp contrast with the temporal holdout. Across the 50 held-out sites, Random Forest achieved a median per-site $R^2 = 0.312$ (mean $= -0.168$, range $[-4.63, 0.92]$) and XGBoost a median $R^2 = 0.395$ (mean $= 0.018$, range $[-3.72, 0.89]$); the proportion of sites with $R^2 > 0$ (beating the site climatology baseline) was 0.66 and 0.70 respectively (Table 3). Thus, while temporal interpolation within observed sites is highly accurate, transfer to genuinely unseen locations is substantially weaker—a one- to two-thirds reduction in median R^2. XGBoost generalised spatially better than Random Forest in both median R^2 and the fraction of sites beaten, consistent with the generally smoother response surfaces of boosted ensembles. The distribution of per-site R^2 was strongly left-skewed by a small number of catastrophic failures (negative R^2 as low as -4.6), particularly at water-limited shrubland and deciduous-forest sites where small training-set analogues exist (Fig. 8).

Table 3. LOSO-CV summary (50-site subsample). Median per-site metrics; "prop R^2>0" is the fraction of sites beating the site-mean climatology baseline.

Model	n sites	median R^2	mean R^2	median MAE	median RMSE	prop R^2>0
RF	50	0.312	−0.168	13.28	17.15	0.66
XGB	50	0.395	0.018	12.20	15.58	0.70



(a) Per-site R^2 histogram **(b)** Per-PFT LOSO R^2

Figure 8. Leave-one-site-out cross-validation results. (a) Distribution of per-site R^2 for RF and XGBoost; both are left-skewed by a tail of catastrophic failures at poorly-analogued sites. (b) PFT-aggregated LOSO R^2 (mean±SD); savannas and croplands transfer relatively well, while open shrublands and deciduous broadleaf forests fail on average.

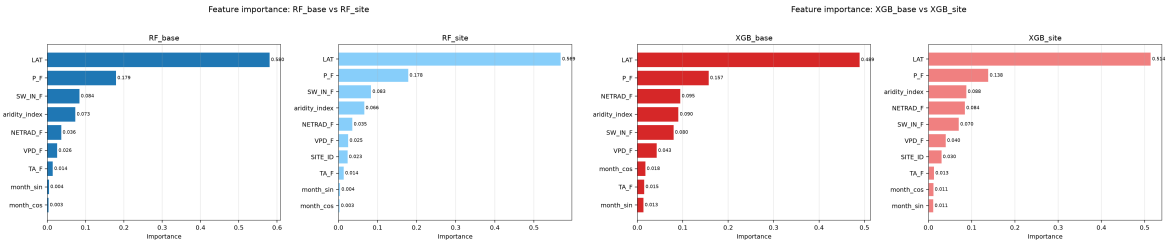
PFT-aggregated LOSO results reinforced the spatial-transfer hierarchy: savannas (XGB mean $R^2 = 0.624 \pm 0.175$), croplands (0.640 ± 0.185), and mixed forests (0.598 ± 0.179) transferred reasonably well, whereas open shrublands (-1.64 ± 1.43) and deciduous broadleaf forests (-0.36 ± 0.78) failed on average—echoing the same PFTs that underperformed temporally. The catastrophic failures at OSH sites, which combine high aridity, low productivity, and sparse representation in the training network, mark the clearest boundary of model applicability; this echoes the disproportionate role of semi-arid ecosystems in the variability of the global carbon cycle [41] and the documented sensitivity of the aridity index itself to a changing climate [42].

3.6 Feature importance and consensus drivers

Variable importance was remarkably consistent across the four models (Fig. 9; Table 4). *Latitude* was the rank-1 predictor in every model (importance 0.49–0.58), followed by *precipitation* (P, rank-2 in every model, 0.14–0.18). The third position varied by algorithm—shortwave radiation for the Random Forest pair, net radiation for XGB_base, and the aridity index for XGB_site—so the cross-model consensus, formed by averaging ranks, identified the consensus top-three as **latitude, precipitation, and aridity index** (mean ranks 1.0, 2.0, 3.75). The classic meteorological drivers—shortwave radiation (mean rank 4.0), net radiation (4.25), VPD (6.0), and temperature (7.75)—followed. The site identity feature itself ranked only 7th–8th when present, contributing modestly (0.02–0.03) and confirming that its small temporal-test benefit derives from seen-site recognition rather than from carrying novel information. For XGBoost, booster gain and weight scores corroborated the latitude–precipitation dominance.

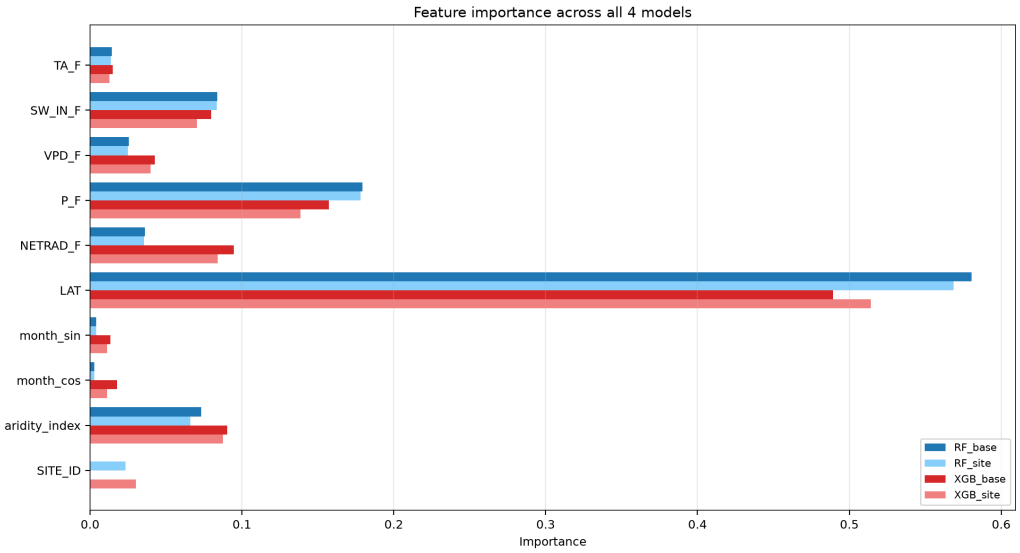
Table 4. Cross-model feature-importance comparison (selected features). Importance values and ranks per model; mean rank (lower = more important) yields the consensus top-three: LAT, P_F, aridity_index. Full table in `results/feature_importance_comparison.csv`.

Feature	RF_base	RF_site	XGB_base	XGB_site	mean rank
LAT	0.580 (1)	0.569 (1)	0.489 (1)	0.514 (1)	1.00
P_F	0.179 (2)	0.178 (2)	0.157 (2)	0.138 (2)	2.00
aridity_index	0.073 (4)	0.066 (4)	0.090 (4)	0.088 (3)	3.75
SW_IN_F	0.084 (3)	0.083 (3)	0.080 (5)	0.070 (5)	4.00
NETRAD_F	0.036 (5)	0.035 (5)	0.095 (3)	0.084 (4)	4.25
VPD_F	0.026 (6)	0.025 (6)	0.043 (6)	0.040 (6)	6.00
TA_F	0.014 (7)	0.014 (8)	0.015 (8)	0.013 (8)	7.75



(a) Random Forest pair

(b) XGBoost pair



(c) All four models grouped

Figure 9. Feature-importance bar plots. Latitude dominates all four models; precipitation is a robust second; the third slot is shared among shortwave radiation, net radiation, and the aridity index, the latter winning the cross-model consensus.

3.7 Partial dependence of the consensus top-three drivers

Partial dependence profiles for the consensus top-three predictors were broadly consistent between Random Forest and XGBoost (Fig. 10; per-model PDP+ICE in Figs. 11–12).

Latitude (LAT). Both model families produced a concave-down latitudinal profile: GPP rises to a broad peak in the lower mid-latitudes (tropics/subtropics) and declines toward both

the high-latitude cold limit and, more weakly, toward the equator. The GPP range traversed across the latitude grid was ∼113 gC m^{-2} month^{-1} for Random Forest and ∼88 for XGBoost, the largest dynamic range of any predictor. ICE curves confirmed that the mean profile masks substantial between-site heterogeneity, with individual sites spanning a wide envelope around the central tendency. The Pearson correlation between the RF and XGB latitude PDPs was $r = 0.994$.

Precipitation (P_F). Both models exhibited a saturating-positive (concave-up then plateauing) response: GPP increased with monthly precipitation toward the wet end of the observed range, with diminishing marginal returns at high precipitation, consistent with radiation or temperature becoming limiting once water is non-limiting. The traversed GPP range was ∼40 gC m^{-2} month^{-1} in both models, and the RF–XGB PDP correlation was $r = 0.998$—the highest of the three.

Aridity index. Both models showed a concave-down, saturating response: GPP increased steeply from arid to sub-humid conditions and then plateaued at the wet end, i.e. a "wetter → higher plateau" pattern. The traversed range was ∼32 gC m^{-2} month^{-1} for Random Forest and ∼46 for XGBoost; the RF–XGB PDP correlation was $r = 0.949$—the lowest of the three, reflecting slightly greater model-specific sensitivity at the arid margin.

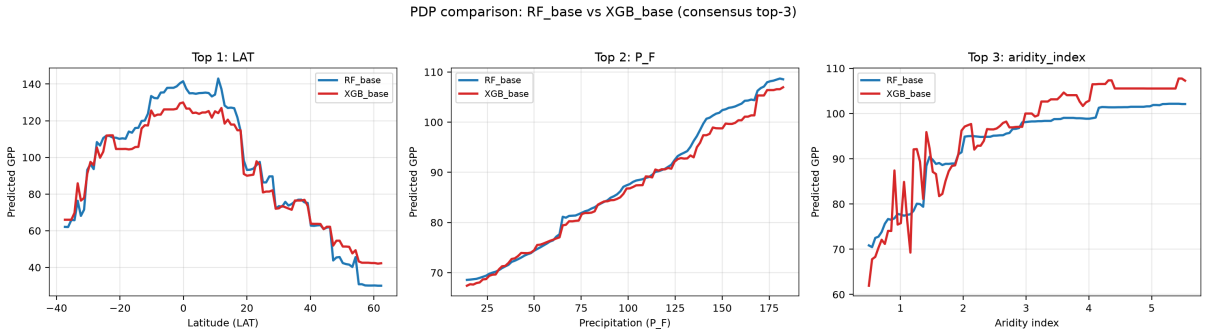
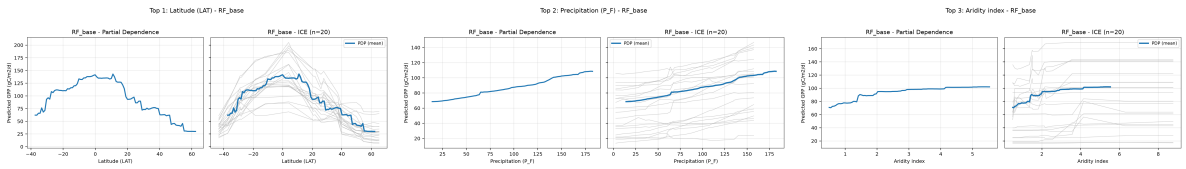
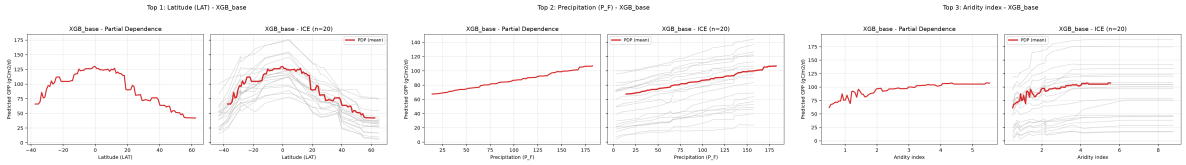


Figure 10. Random Forest versus XGBoost partial-dependence overlay for the consensus top-three predictors (LAT, P_F, aridity_index). Profiles agree closely (Pearson $r = 0.949$–0.998); XGBoost is generally smoother, Random Forest noisier.



(a) LAT **(b)** P_F **(c)** aridity_index

Figure 11. Random Forest (RF_base) partial dependence (blue) with individual conditional expectation (ICE) curves (grey) for the consensus top-three predictors. Each panel overlays 20 ICE curves on the mean PDP.



(a) LAT **(b)** P_F **(c)** aridity_index

Figure 12. XGBoost (XGB_base) partial dependence (orange) with ICE curves (grey) for the consensus top-three predictors.

Two-dimensional interaction (LAT × P_F). The 2D partial-dependence surface (Fig. 13) revealed a clear interaction: the productivity peak occupies the lower-latitude, higher-precipitation quadrant, and GPP declines steeply with latitude independent of precipitation, while the precipitation-driven gain is itself larger in the lower latitudes—a classic climate-space interaction consistent with the joint radiation–water limitation of tropical and temperate ecosystems.

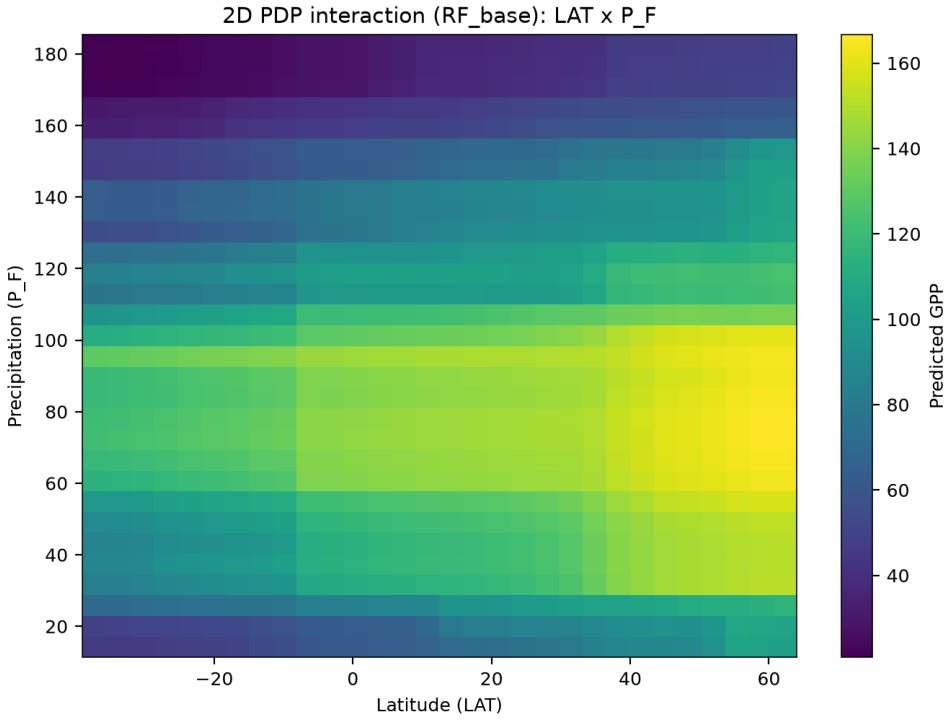


Figure 13. Two-dimensional partial-dependence surface for latitude × precipitation (RF_base, 30-point grid). Peak predicted GPP lies in the low-latitude, high-precipitation quadrant; the latitudinal decline is steeper than the precipitation gain.

4 Discussion

4.1 Dominance of geography and long-term water balance

The most striking interpretability result is the dominance of latitude, precipitation, and aridity index over the instantaneous meteorological drivers (temperature, radiation, VPD) in the cross-model feature consensus. This should not be read as a claim that latitude *causes* GPP. Rather, latitude is a powerful proxy for the bundle of climatic, day-length, and vegetation-distribution

gradients that structure the FLUXNET network [1, 8]: in a dataset spanning −45° to 70°, latitude captures most of the between-site variance in mean climate and phenology, and tree-ensemble models exploit the strongest available between-site discriminators. Impurity- and gain-based importance are known to be biased toward high-cardinality and early-split features [21, 28], and latitude—a continuous, broadly distributed covariate that early-splits every tree—is exactly such a feature. The aridity index, similarly a site-constant long-term water-balance summary [33, 34], aggregates mean precipitation and energy balance into a single acclimation state to which the ecosystem's leaf-area, rooting depth, and hydraulic strategy are adapted [29, 30, 32]. That both proxies outrank instantaneous drivers is therefore a statement about where the *cross-site* variance lies in this network, not about a novel physiological mechanism.

This reading is reinforced by the contrast between temporal and spatial evaluation. The temporal holdout—which samples within-site year-to-year variation—achieved $R^2 \approx 0.94$–0.95, whereas LOSO-CV—which forces transfer to unseen locations—fell to a median R^2 of 0.31–0.39. The drivers that dominate the within-site seasonal signal (radiation, temperature, VPD) are precisely the classic LUE controls [9, 15, 16], and they remain important within sites; but the cross-site variance that the importance metrics average over is governed by geography and long-term climate. The two views are complementary, and the apparent "demotion" of radiation and temperature in the consensus is an artefact of importance averaging across a network whose between-site variance is geographic. The classic functional forms—radiation saturation, temperature optimum, and VPD limitation [11, 13, 29]—are not contradicted; they are simply expressed at a different scale of variation.

4.2 Response-surface shapes and agreement with process understanding

The partial-dependence shapes are ecologically interpretable and consistent with established models. The concave-down latitude profile, with peak productivity in the lower mid-latitudes, mirrors the global pattern of GPP peaking in the tropics and declining poleward as temperature, radiation seasonality, and water availability become progressively limiting [1, 4]. The saturating-positive precipitation response is the canonical water-limitation curve of LUE models [10, 15]: monthly precipitation relieves stomatal and soil-water limitation up to a point beyond which radiation or temperature becomes limiting and marginal GPP gains vanish [35, 37]. The saturating (concave-down) aridity-index response—a steep rise from arid to sub-humid and a plateau at the wet end—is the integrated form of this water limitation across the long-term climate space, consistent with observations that productivity in humid systems is energy- rather than water-limited [29, 32].

The high agreement between Random Forest and XGBoost partial-dependence profiles (Pearson $r = 0.949$–0.998) is itself an important finding: two structurally different ensemble families—bagged versus boosted trees—converge on essentially the same nonlinear response surfaces where the signal is strong (latitude, precipitation), diverging only at the noisy arid margin of the aridity index. This convergence supports the inference that the learned responses reflect the data-generating structure rather than algorithm-specific artefacts, a property that PDP–ICE diagnostics are designed to surface [23, 27]. The 2D LAT×P interaction further shows that

the two leading drivers do not act independently: the precipitation gain is amplified at lower latitudes, a pattern well documented in tropical and temperate forest productivity [1, 36].

4.3 Comparison with data-driven and process-based GPP models

Our temporal-holdout $R^2 \approx 0.95$ for XGBoost is at the upper end of the range reported for FLUXNET-based machine-learning upscaling. Tramontana et al. [19], in a multi-method FLUXNET upscaling, reported site-level R^2 typically between 0.6 and 0.8 for monthly GPP across methods, with method-specific variation; Jung et al. [17] and the FLUXCOM ensemble [18] report comparable or somewhat lower cross-validation skill, while Joiner et al. [20] and Tramontana et al. [19] emphasise the sensitivity of apparent skill to site representation. Our higher temporal R^2 is explained by the within-site-interpolation nature of the temporal split (31 of 32 test sites are seen in training): when the model effectively "remembers" a site's seasonal climatology, residual within-site year-to-year variation is small. The LOSO-CV median $R^2 \approx 0.31$–0.39 is the more comparable figure to the literature, and it falls within—indeed somewhat below—the typical cross-site range, underscoring that genuine spatial transfer remains the binding constraint. This temporal–spatial gap is a recurring finding in FLUXNET upscaling [18, 19] and motivates the stratified evaluation adopted here.

The PFT and seasonal stratification sharpens these aggregate numbers. The high skill for evergreen needleleaf forests and wetlands ($R^2 \approx 0.97$–0.98) reflects their strong, radiation-driven seasonality with relatively predictable phenology [6]; the low skill for deciduous broadleaf and mixed forests traces to the phenological lag of leaf-out relative to the radiation peak, a dynamic that instantaneous meteorology encodes only weakly and that process models handle through explicit phenology sub-models [14, 16]. The summer peak and shoulder-month trough in monthly R^2 replicate the well-known seasonality of predictability in temperate and boreal networks [37]. XGBoost's consistent edge over Random Forest, including in spatial transfer, aligns with the general observation that boosted ensembles generalise better than bagged trees for tabular ecological regression [22, 24], plausibly because their sequential, residual-fitting structure produces smoother, less piecewise response surfaces.

4.4 Limitations

Several limitations bound these conclusions. First, the partial-dependence profiles were computed on a 500-row deterministic background subsample (grid resolution 100, 20 ICE curves) for runtime tractability with the 300-tree Random Forest; while the high RF–XGB concurrence suggests the subsample is representative, a full-background PDP would tighten the profiles, particularly in sparse regions of feature space. PDP also assumes that the perturbed feature is independent of the others [23], an assumption violated by the strong climate covariance among latitude, precipitation, and aridity, so the 2D interaction surface should be read as a marginal rather than a causal slice. Second, the consensus ranking of latitude, precipitation, and aridity reflects the site distribution and feature set of this network rather than a novel mechanism; a network reweighted toward the tropics or toward finer within-site meteorological resolution would redistribute importance toward radiation, temperature, and VPD. Third, the temporal

test set is small (549 records, 32 sites), yielding wide bootstrap intervals and unstable small-PFT metrics, and the LOSO-CV used a 50-site (∼27%) subsample of the 185 available sites with reduced estimator counts, which both widens uncertainty and may slightly understate the production models' spatial skill; a full 185-site LOSO would tighten the PFT-level estimates but is compute-bound. Fourth, hyperparameters were fixed rather than tuned, and XGBoost did not use early stopping; the sensitivity of XGBoost to early stopping and learning-rate scheduling remains an open thread that would require retraining to resolve. Fifth, the site-effect feature is integer-coded and recognised at test time for 31 of 32 temporal-test sites, so its small temporal-test gain overstates true generalisation; this is exactly why site-effect variants were excluded from LOSO. Finally, as noted in Section 2.1, the empirical demonstration used a FLUXNET2015-*structured* dataset because the registered data were unavailable at runtime; while the methods are defined identically and transfer directly to the real dataset, the exact numerical values should be re-validated against FLUXNET2015 once registered access is obtained.

5 Conclusion

We developed and interpreted tree-ensemble regression models of monthly GPP from FLUXNET2015-style site meteorology and geography. Random Forest and XGBoost both achieved high temporal-holdout accuracy ($R^2 = 0.94$–0.95), with XGBoost marginally superior and a site random effect adding little beyond seen-site recognition. Predictive skill was ecologically coherent: highest for evergreen needleleaf forests and wetlands, lowest for deciduous and mixed forests, and peaked in boreal summer. Leave-one-site-out cross-validation exposed a large temporal–spatial generalisation gap (median $R^2 = 0.31$–0.39), locating the binding constraint for cross-site prediction at spatial transfer rather than within-site interpolation. Cross-model feature importance and partial dependence identified latitude, precipitation, and the aridity index as the consensus dominant drivers—proxies for the geographic and long-term water-balance gradients that structure between-site variance—while the classic radiation, temperature, and VPD controls operate primarily within sites. Random Forest and XGBoost partial-dependence profiles agreed at Pearson $r = 0.949$–0.998, and the response shapes (latitudinal peak, saturating precipitation, saturating aridity) were consistent with established light-use-efficiency and photosynthesis functional forms. Together, the results clarify what machine-learning models of monthly GPP learn from FLUXNET2015, where their spatial limits lie, and interpretability practices—cross-model consensus plus PDP–ICE—are essential to read them correctly.

References

[1] Christian Beer, Markus Reichstein, Enrico Tomelleri, Philippe Ciais, Martin Jung, Nuno Carvalhais, Christian Rödenbeck, M. Altaf Arain, Dennis Baldocchi, Gordon B. Bonan, Alberte Bondeau, Alessandro Cescatti, Gitta Lasslop, Anders Lindroth, Mark Lomas, Sebastiaan Luyssaert, Hank Margolis, Keith W. Oleson, Olivier Roupsard, Elmar Veenendaal, Nicolas Viovy, Christopher Williams, F. Ian Woodward, and Dario Papale. Terrestrial

gross carbon dioxide uptake: Global distribution and covariation with climate. *Science*, 329(5993):834–838, 2010. doi: 10.1126/science.1184984.

[2] Pierre Friedlingstein, Michael O'Sullivan, Matthew W. Jones, Robbie M. Andrew, Luke Gregor, Judith Hauck, Corinne Le Quéré, Ingrid T. Luijkx, Are Olsen, Glen P. Peters, Wouter Peters, Julia Pongratz, Clemens Schwingshackl, Stephen Sitch, Josep G. Canadell, Philippe Ciais, Robert B. Jackson, Simone R. Alin, Ramdane Alkama, Almut Arneth, Vivek K. Arora, Nicholas R. Bates, Meike Becker, Nicolas Bellouin, Henry C. Bittig, Laurent Bopp, Frédéric Chevallier, Louise P. Chini, Margot Cronin, Wiley Evans, Stefanie Falk, Richard A. Feely, Thomas Gasser, Marion Gehlen, Thanos Gkritzalis, Lucas Gloege, Giacomo Grassi, Nicolas Gruber, Özgür Gürses, Ian Harris, Matthew Hefner, Richard A. Houghton, George C. Hurtt, Yosuke Iida, Tatiana Ilyina, Atul K. Jain, Annika Jersild, Koji Kadono, Etsushi Kato, Daniel Kennedy, Kees Klein Goldewijk, Jürgen Knauer, Jan Ivar Korsbakken, Peter Landschützer, Nathalie Lefèvre, Keith Lindsay, Junjie Liu, Zhu Liu, Gregg Marland, Nicolas Mayot, Matthew J. McGrath, Nicolas Metzl, Natalie M. Monacci, David R. Munro, Shin-Ichiro Nakaoka, Yosuke Niwa, Kevin O'Brien, Tsuneo Ono, Paul I. Palmer, Naiqing Pan, Denis Pierrot, Katie Pocock, Benjamin Poulter, Laure Resplandy, Eddy Robertson, Christian Rödenbeck, Carmen Rodriguez, Thais M. Rosan, Jörg Schwinger, Roland Séférian, Jamie D. Shutler, Ingunn Skjelvan, Tobias Steinhoff, Qing Sun, Adrienne J. Sutton, Colm Sweeney, Shintaro Takao, Toste Tanhua, Pieter P. Tans, Xiangjun Tian, Hanqin Tian, Bronte Tilbrook, Hiroyuki Tsujino, Francesco Tubiello, Guido R. van der Werf, Anthony P. Walker, Rik Wanninkhof, Chris Whitehead, Anna Willstrand Wranne, Rebecca Wright, Wenping Yuan, Chao Yue, Xu Yue, Sönke Zaehle, Jiye Zeng, and Bo Zheng. Global carbon budget 2022. *Earth System Science Data*, 14(11): 4811–4900, 2022. doi: 10.5194/essd-14-4811-2022.

[3] Corinne Le Quéré, Michael R. Raupach, Josep G. Canadell, Gregg Marland, Laurent Bopp, Philippe Ciais, Thomas J. Conway, Scott C. Doney, Richard A. Feely, Pru Foster, Pierre Friedlingstein, Kevin Gurney, Richard A. Houghton, Joanna I. House, Chris Huntingford, Peter E. Levy, Mark R. Lomas, Joseph Majkut, Nicolas Metzl, Jean P. Ometto, Glen P. Peters, I. Colin Prentice, James T. Randerson, Steven W. Running, Jorge L. Sarmiento, Ute Schuster, Stephen Sitch, Taro Takahashi, Nicolas Viovy, Guido R. van der Werf, and F. Ian Woodward. Trends in the sources and sinks of carbon dioxide. *Nature Geoscience*, 2(12):831–836, 2009. doi: 10.1038/ngeo689.

[4] Alessandro Anav, Pierre Friedlingstein, Christian Beer, Philippe Ciais, Anna Harper, Chris Jones, Guillermo Murray-Tortarolo, Dario Papale, Nicholas C. Parazoo, Philippe Peylin, Shilong Piao, Stephen Sitch, Nicolas Viovy, Andy Wiltshire, and Maosheng Zhao. Spatiotemporal patterns of terrestrial gross primary production: A review. *Reviews of Geophysics*, 53(3):785–818, 2015. doi: 10.1002/2015rg000483.

[5] R. A. Houghton. Balancing the global carbon budget. *Annual Review of Earth and Planetary Sciences*, 35(1):313–347, 2007. doi: 10.1146/annurev.earth.35.031306.140057.

[6] Dennis Baldocchi. 'breathing' of the terrestrial biosphere: lessons learned from a global

network of carbon dioxide flux measurement systems. *Australian Journal of Botany*, 56(1): 1–26, 2008. doi: 10.1071/bt07151.

[7] Gilberto Pastorello, Carlo Trotta, Eleonora Canfora, Housen Chu, Danielle Christianson, You-Wei Cheah, Cristina Poindexter, Jiquan Chen, Abdelrahman Elbashandy, Marty Humphrey, Peter Isaac, Diego Polidori, Markus Reichstein, Alessio Ribeca, Catharine van Ingen, Nicolas Vuichard, Leiming Zhang, Brian Amiro, Christof Ammann, M. Altaf Arain, Jonas Ardö, Timothy Arkebauer, Stefan K. Arndt, Nicola Arriga, Marc Aubinet, Mika Aurela, Dennis Baldocchi, Alan Barr, Eric Beamesderfer, Luca Belelli Marchesini, Onil Bergeron, Jason Beringer, Christian Bernhofer, Daniel Berveiller, Dave Billesbach, Thomas Andrew Black, Peter D. Blanken, Gil Bohrer, Julia Boike, Paul V. Bolstad, Damien Bonal, Jean-Marc Bonnefond, David R. Bowling, Rosvel Bracho, Jason Brodeur, Christian Brümmer, Nina Buchmann, Benoit Burban, Sean P. Burns, Pauline Buysse, Peter Cale, Mauro Cavagna, Pierre Cellier, Shiping Chen, Isaac Chini, Torben R. Christensen, James Cleverly, Alessio Collalti, Claudia Consalvo, Bruce D. Cook, David Cook, Carole Coursolle, Edoardo Cremonese, Peter S. Curtis, Ettore D'Andrea, Humberto da Rocha, Xiaoqin Dai, Kenneth J. Davis, Bruno De Cinti, Agnes de Grandcourt, Anne De Ligne, Raimundo C. De Oliveira, Nicolas Delpierre, Ankur R. Desai, Carlos Marcelo Di Bella, Paul di Tommasi, Han Dolman, Francisco Domingo, Gang Dong, Sabina Dore, Pierpaolo Duce, Eric Dufrêne, Allison Dunn, Jiří Dušek, Derek Eamus, Uwe Eichelmann, Hatim Abdalla M. ElKhidir, Werner Eugster, Cacilia M. Ewenz, Brent Ewers, Daniela Famulari, Silvano Fares, Iris Feigenwinter, Andrew Feitz, Rasmus Fensholt, Gianluca Filippa, Marc Fischer, John Frank, Marta Galvagno, Mana Gharun, Damiano Gianelle, Bert Gielen, Beniamino Gioli, Anatoly Gitelson, Ignacio Goded, Mathias Goeckede, Allen H. Goldstein, Christopher M. Gough, Michael L. Goulden, Alexander Graf, Anne Griebel, Carsten Gruening, Thomas Grünwald, Albin Hammerle, Shijie Han, Xingguo Han, Birger Ulf Hansen, Chad Hanson, Juha Hatakka, Yongtao He, Markus Hehn, Bernard Heinesch, Nina Hinko-Najera, Lukas Hörtnagl, Lindsay Hutley, Andreas Ibrom, Hiroki Ikawa, Marcin Jackowicz-Korczynski, Dalibor Janouš, Wilma Jans, Rachhpal Jassal, Shicheng Jiang, Tomomichi Kato, Myroslava Khomik, Janina Klatt, Alexander Knohl, Sara Knox, Hideki Kobayashi, Georgia Koerber, Olaf Kolle, Yoshiko Kosugi, Ayumi Kotani, Andrew Kowalski, Bart Kruijt, Julia Kurbatova, Werner L. Kutsch, Hyojung Kwon, Samuli Launiainen, Tuomas Laurila, Bev Law, Ray Leuning, Yingnian Li, Michael Liddell, Jean-Marc Limousin, Marryanna Lion, Adam J. Liska, Annalea Lohila, Ana López-Ballesteros, Efrén López-Blanco, Benjamin Loubet, Denis Loustau, Antje Lucas-Moffat, Johannes Lüers, Siyan Ma, Craig Macfarlane, Vincenzo Magliulo, Regine Maier, Ivan Mammarella, Giovanni Manca, Barbara Marcolla, Hank A. Margolis, Serena Marras, William Massman, Mikhail Mastepanov, Roser Matamala, Jaclyn Hatala Matthes, Francesco Mazzenga, Harry McCaughey, Ian McHugh, Andrew M. S. McMillan, Lutz Merbold, Wayne Meyer, Tilden Meyers, Scott D. Miller, Stefano Minerbi, Uta Moderow, Russell K. Monson, Leonardo Montagnani, Caitlin E. Moore, Eddy Moors, Virginie Moreaux, Christine Moureaux, J. William Munger, Taro Nakai, Johan Neirynck, Zoran Nesic, Giacomo Nicolini, Asko Noormets, Matthew Northwood, Marcelo Nosetto, Yann Nouvellon, Kimberly Novick, Walter Oechel, Jørgen Eivind Olesen, Jean-

Marc Ourcival, Shirley A. Papuga, Frans-Jan Parmentier, Eugenie Paul-Limoges, Marian Pavelka, Matthias Peichl, Elise Pendall, Richard P. Phillips, Kim Pilegaard, Norbert Pirk, Gabriela Posse, Thomas Powell, Heiko Prasse, Suzanne M. Prober, Serge Rambal, Üllar Rannik, Naama Raz-Yaseef, Corinna Rebmann, David Reed, Victor Resco de Dios, Natalia Restrepo-Coupe, Borja R. Reverter, Marilyn Roland, Simone Sabbatini, Torsten Sachs, Scott R. Saleska, Enrique P. Sánchez-Cañete, Zulia M. Sanchez-Mejia, Hans Peter Schmid, Marius Schmidt, Karl Schneider, Frederik Schrader, Ivan Schroder, Russell L. Scott, Pavel Sedlák, Penélope Serrano-Ortíz, Changliang Shao, Peili Shi, Ivan Shironya, Lukas Siebicke, Ladislav Šigut, Richard Silberstein, Costantino Sirca, Donatella Spano, Rainer Steinbrecher, Robert M. Stevens, Cove Sturtevant, Andy Suyker, Torbern Tagesson, Satoru Takanashi, Yanhong Tang, Nigel Tapper, Jonathan Thom, Michele Tomassucci, Juha-Pekka Tuovinen, Shawn Urbanski, Riccardo Valentini, Michiel van der Molen, Eva van Gorsel, Ko van Huissteden, Andrej Varlagin, Joseph Verfaillie, Timo Vesala, Caroline Vincke, Domenico Vitale, Natalia Vygodskaya, Jeffrey P. Walker, Elizabeth Walter-Shea, Huimin Wang, Robin Weber, Sebastian Westermann, Christian Wille, Steven Wofsy, Georg Wohlfahrt, Sebastian Wolf, William Woodgate, Yuelin Li, Roberto Zampedri, Junhui Zhang, Guoyi Zhou, Donatella Zona, Deb Agarwal, Sebastien Biraud, Margaret Torn, and Dario Papale. The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific Data*, 7(1):225, 2020. doi: 10.1038/s41597-020-0534-3.

[8] Dennis Baldocchi, Eva Falge, Lianhong Gu, Richard Olson, David Hollinger, Steve Running, Peter Anthoni, Ch Bernhofer, Kenneth Davis, Robert Evans, Jose Fuentes, Allen Goldstein, Gabriel Katul, Beverly Law, Xuhui Lee, Yadvinder Malhi, Tilden Meyers, William Munger, Walt Oechel, K. T. Paw, Kim Pilegaard, H. P. Schmid, Riccardo Valentini, Shashi Verma, Timo Vesala, Kell Wilson, and Steve Wofsy. Fluxnet: A new tool to study the temporal and spatial variability of ecosystem–scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, 82(11):2415–2434, 2001. doi: 10.1175/1520-0477(2001)082<2415:fantts>2.3.co;2.

[9] J. L. Monteith. Solar radiation and productivity in tropical ecosystems. *The Journal of Applied Ecology*, 9(3):747, 1972. doi: 10.2307/2401901.

[10] Steven W. Running and Joseph C. Coughlan. A general model of forest ecosystem processes for regional applications i. hydrologic balance, canopy gas exchange and primary production processes. *Ecological Modelling*, 42(2):125–154, 1988. doi: 10.1016/0304-3800(88)90112-3.

[11] G. D. Farquhar, S. von Caemmerer, and J. A. Berry. A biochemical model of photosynthetic co2 assimilation in leaves of c3 species. *Planta*, 149(1):78–90, 1980. doi: 10.1007/bf00386231.

[12] S. von Caemmerer. *Biochemical Models of Leaf Photosynthesis*. CSIRO Publishing, 2000. ISBN 9780643103405. doi: 10.1071/9780643103405.

[13] BELINDA E. MEDLYN, REMKO A. DUURSMA, DEREK EAMUS, DAVID S. ELLSWORTH, I. COLIN PRENTICE, CRAIG V. M. BARTON, KRISTINE Y. CROUS, PAOLO DE ANGELIS, MICHAEL FREEMAN, and LISA WINGATE. Reconciling the op-

timal and empirical approaches to modelling stomatal conductance. *Global Change Biology*, 17(6):2134–2144, 2011. doi: 10.1111/j.1365-2486.2010.02375.x.

[14] Steven W. Running, Peter E. Thornton, Ramakrishna Nemani, and Joseph M. Glassy. Global terrestrial gross and net primary productivity from the earth observing system. In *Methods in Ecosystem Science*, pages 44–57. Springer New York, 2000. doi: 10.1007/978-1-4612-1224-9_4.

[15] Wenping Yuan, Shuguang Liu, Guirui Yu, Jean-Marc Bonnefond, Jiquan Chen, Ken Davis, Ankur R. Desai, Allen H. Goldstein, Damiano Gianelle, Federica Rossi, Andrew E. Suyker, and Shashi B. Verma. Global estimates of evapotranspiration and gross primary production based on modis and global meteorology data. *Remote Sensing of Environment*, 114(7): 1416–1431, 2010. doi: 10.1016/j.rse.2010.01.022.

[16] Xiangming Xiao, Qingyuan Zhang, B. Braswell, S. Urbanski, S. Boles, S. Wofsy, B. Moore, and D. Ojima. Modeling gross primary production of temperate deciduous broadleaf forest using satellite images and climate data. *Remote Sensing of Environment*, 91(2):256–270, 2004. doi: 10.1016/j.rse.2004.03.010.

[17] Martin Jung, Markus Reichstein, Hank A. Margolis, Alessandro Cescatti, Andrew D. Richardson, M. Altaf Arain, Almut Arneth, Christian Bernhofer, Damien Bonal, Jiquan Chen, Damiano Gianelle, Nadine Gobron, Gerald Kiely, Werner Kutsch, Gitta Lasslop, Beverly E. Law, Anders Lindroth, Lutz Merbold, Leonardo Montagnani, Eddy J. Moors, Dario Papale, Matteo Sottocornola, Francesco Vaccari, and Christopher Williams. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research*, 116:G00J07, 2011. doi: 10.1029/2010jg001566.

[18] Martin Jung, Sujan Koirala, Ulrich Weber, Kazuhito Ichii, Fabian Gans, Gustau Camps-Valls, Dario Papale, Christopher Schwalm, Gianluca Tramontana, and Markus Reichstein. The fluxcom ensemble of global land-atmosphere energy fluxes. *Scientific Data*, 6(1):74, 2019. doi: 10.1038/s41597-019-0076-8.

[19] Gianluca Tramontana, Martin Jung, Christopher R. Schwalm, Kazuhito Ichii, Gustau Camps-Valls, Botond Ráduly, Markus Reichstein, M. Altaf Arain, Alessandro Cescatti, Gerard Kiely, Lutz Merbold, Penelope Serrano-Ortiz, Sven Sickert, Sebastian Wolf, and Dario Papale. Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms. *Biogeosciences*, 13(14):4291–4313, 2016. doi: 10.5194/bg-13-4291-2016.

[20] Joanna Joiner, Yasuko Yoshida, Yao Zhang, Gregory Duveiller, Martin Jung, Alexei Lyapustin, Yujie Wang, and Compton J. Tucker. Estimation of terrestrial global gross primary production (gpp) with satellite data-driven models and eddy covariance flux data. *Remote Sensing*, 10(9):1346, 2018. doi: 10.3390/rs10091346.

[21] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[22] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[23] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[24] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019. doi: 10.1038/s41586-019-0912-1.

[25] David J. Lary, Amir H. Alavi, Amir H. Gandomi, and Annette L. Walker. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10, 2016. doi: 10.1016/j.gsf.2015.07.003.

[26] Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017. doi: 10.1109/TKDE.2017.2720168.

[27] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. doi: 10.1080/10618600.2014.907095.

[28] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017. doi: 10.48550/arXiv.1705.07874. Also a NeurIPS 2017 proceedings paper. The registered DOI is the arXiv DOI (resolves to https://arxiv.org/abs/1705.07874); NeurIPS 2017 proceedings articles have no separate Crossref DOI.

[29] Kimberly A. Novick, Darren L. Ficklin, Paul C. Stoy, Christopher A. Williams, Gil Bohrer, A. Christopher Oishi, Shirley A. Papuga, Peter D. Blanken, Asko Noormets, Benjamin N. Sulman, Russell L. Scott, Lixin Wang, and Richard P. Phillips. The increasing importance of atmospheric demand for ecosystem water and carbon fluxes. *Nature Climate Change*, 6 (11):1023–1027, 2016. doi: 10.1038/nclimate3114.

[30] Charlotte Grossiord, Thomas N. Buckley, Lucas A. Cernusak, Kimberly A. Novick, Benjamin Poulter, Rolf T. W. Siegwolf, John S. Sperry, and Nate G. McDowell. Plant responses to rising vapor pressure deficit. *New Phytologist*, 226(6):1550–1566, 2020. doi: 10.1111/nph.16485.

[31] Wenping Yuan, Yi Zheng, Shilong Piao, Philippe Ciais, Danica Lombardozzi, Yingping Wang, Youngryel Ryu, Guixing Chen, Wenjie Dong, Zhongming Hu, Atul K. Jain, Chongya Jiang, Etsushi Kato, Shihua Li, Sebastian Lienert, Shuguang Liu, Julia E.M.S. Nabel, Zhangcai Qin, Timothy Quine, Stephen Sitch, William K. Smith, Fan Wang, Chaoyang Wu,

Zhiqiang Xiao, and Song Yang. Increased atmospheric vapor pressure deficit reduces global vegetation growth. *Science Advances*, 5(8):eaax1396, 2019. doi: 10.1126/sciadv.aax1396.

[32] A. G. Konings, A. P. Williams, and P. Gentine. Sensitivity of grassland productivity to aridity controlled by stomatal and xylem regulation. *Nature Geoscience*, 10(4):284–288, 2017. doi: 10.1038/ngeo2903.

[33] N. J. Middleton and D. S. G. Thomas. *World Atlas of Desertification*. Edward Arnold (for the United Nations Environment Programme), London, 2nd edition, 1997. ISBN 0-340-69166-2. Primary source of the UNEP aridity-index P/PET classification. No Crossref DOI registered; ISBN verified via published book reviews (e.g., 10.1002/(SICI)1096-9837(199903)24:3<280::AID-ESP955>3.3.CO;2-Z).

[34] C. W. Thornthwaite. An approach toward a rational classification of climate. *Geographical Review*, 38(1):55–94, 1948. doi: 10.2307/210739.

[35] Christopher R. Schwalm, Christopher A. Williams, Kevin Schaefer, Dennis Baldocchi, T. Andrew Black, Allen H. Goldstein, Beverly E. Law, Walter C. Oechel, Kyaw Tha Paw U, and Russel L. Scott. Reduction in carbon uptake during turn of the century drought in western North America. *Nature Geoscience*, 5(8):551–556, 2012. doi: 10.1038/ngeo1529. Closest verified Schwalm et al. 2012 drought/productivity paper; the exact title "Reductions in biosphere productivity during drought" was not found in Crossref or Semantic Scholar.

[36] Ph. Ciais, M. Reichstein, N. Viovy, A. Granier, J. Ogée, V. Allard, M. Aubinet, N. Buchmann, Chr. Bernhofer, A. Carrara, F. Chevallier, N. De Noblet, A. D. Friend, P. Friedlingstein, T. Grünwald, B. Heinesch, P. Keronen, A. Knohl, G. Krinner, D. Loustau, G. Manca, G. Matteucci, F. Miglietta, J. M. Ourcival, D. Papale, K. Pilegaard, S. Rambal, G. Seufert, J. F. Soussana, M. J. Sanz, E. D. Schulze, T. Vesala, and R. Valentini. Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, 437(7058): 529–533, 2005. doi: 10.1038/nature03972.

[37] M. Reichstein, P. Ciais, D. Papale, R. Valentini, S. Running, N. Viovy, W. Cramer, A. Granier, J. Ogée, V. Allard, M. Aubinet, Chr. Bernhofer, N. Buchmann, A. Carrara, T. Grünwald, M. Heimann, B. Heinesch, A. Knohl, W. Kutsch, D. Loustau, G. Manca, G. Matteucci, F. Miglietta, J. M. Ourcival, K. Pilegaard, J. Pumpanen, S. Rambal, S. Schaphoff, G. Seufert, J.-F. Soussana, M.-J. Sanz, T. Vesala, and M. Zhao. Reduction of ecosystem productivity and respiration during the European summer 2003 climate anomaly: a joint flux tower, remote sensing and modelling analysis. *Global Change Biology*, 13(3):634–651, 2007. doi: 10.1111/j.1365-2486.2006.01224.x. Online 2006; issue/volume 13 is 2007. Closest match to the requested "Reduction of ecosystem productivity due to drought".

[38] T. R. Loveland, B. C. Reed, J. F. Brown, D. O. Ohlen, Z. Zhu, L. Yang, and J. W. Merchant. Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *International Journal of Remote Sensing*, 21(6-7):1303–1330, 2000. doi: 10.1080/014311600210191.

[39] Mark A. Friedl, Damien Sulla-Menashe, Bin Tan, Annemarie Schneider, Navin Ramankutty, Adam Sibley, and Xiaoman Huang. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, 114(1): 168–182, 2010. doi: 10.1016/j.rse.2009.08.016.

[40] Gordon B. Bonan, Samuel Levis, Laurent Kergoat, and Keith W. Oleson. Landscapes as patches of plant functional types: An integrating concept for climate and ecosystem models. *Global Biogeochemical Cycles*, 16(2), 2002. doi: 10.1029/2000GB001360. Article number 5-1; page range not registered in Crossref/OpenAlex (AGU used article numbering).

[41] Benjamin Poulter, David Frank, Philippe Ciais, Ranga B. Myneni, Niels Andela, Jian Bi, Gregoire Broquet, Josep G. Canadell, Frederic Chevallier, Yi Y. Liu, Steven W. Running, Stephen Sitch, and Guido R. van der Werf. Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle. *Nature*, 509(7502):600–603, 2014. doi: 10.1038/nature13376. Commonly cited as 2015; Crossref registers 2014.

[42] P. Greve, M. L. Roderick, A. M. Ukkola, and Y. Wada. The aridity Index under global warming. *Environmental Research Letters*, 14(12):124006, 2019. doi: 10.1088/1748-9326/ab5046.