FLUXPART-E: A BENCHMARK FOR EVALUATING CO_2 FLUX PARTITIONING METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating whether land ecosystems will keep absorbing part of our CO_2 emissions or begin releasing carbon requires separating two opposing fluxes—photosynthetic uptake (GPP) and ecosystem respiration ($\mathrm{R_{eco}}$)—that are never measured directly: eddy covariance, the most prominent and least invasive method, observes only their net signal, the net ecosystem exchange (NEE). *CO_2 flux partitioning* recovers GPP and $\mathrm{R_{eco}}$ from NEE and meteorological drivers under $\mathrm{NEE} = -\mathrm{GPP} + \mathrm{R_{eco}}$. It is an under-determined inverse problem: one constraint, two unknowns, closed only by assumptions that trade flexibility for identifiability, with noise, gaps, and abrupt events corrupting what little signal remains. It also inverts the usual supervised setting: the target is never observed, so a method is applied to the very data it fits. We introduce **FluxPart-E**, a benchmark and evaluation protocol that makes the problem measurable. Ground truth comes from two structurally distinct land surface models driven by meteorology at 188 observational sites, into which we reinject measurement-realistic noise, gaps, and dynamic rain pulse events. The protocol scores the latent decomposition into GPP and $\mathrm{R_{eco}}$ across defined stress regimes. FluxPart-E exposes failure modes—low data coverage, the edges of climate zones, peak temperatures—that leave detectable signatures in real eddy-covariance data, turning a domain problem into a concrete machine-learning challenge: injecting knowledge, transfer across space and time, and learning reliably under sparse, noisy, and shifting conditions to solve a real inverse problem.

1 INTRODUCTION

Will the terrestrial biosphere continue to act as a carbon sink, or turn into a source? Answering this central question of carbon-cycle science (Baldocchi, 2020) requires more than the net carbon exchange between land and atmosphere: it requires resolving the two opposing processes behind it, gross primary production (GPP), the photosynthetic uptake of CO_2, and ecosystem respiration ($\mathrm{R_{eco}}$), its respiratory release. These component fluxes, and not net ecosystem exchange (NEE) alone, are what let the field ask where and why the land sink is at risk (Musavi et al., 2017; Mahecha et al., 2010). While NEE is measured directly at sites worldwide with the eddy covariance (EC) technique (Baldocchi et al., 2001), GPP and $\mathrm{R_{eco}}$ are not: they must be inferred by *CO_2 flux partitioning*, so every component-level result rests on a partitioning step. The task looks simple: given a noisy $\mathrm{NEE_{obs}} = \mathrm{NEE} + \varepsilon$, with measurement noise ε, recover the GPP and $\mathrm{R_{eco}}$ components whose sum reproduces the underlying, noise-free flux,

$$\mathrm{NEE_{obs}} = -\mathrm{GPP} + \mathrm{R_{eco}} + \varepsilon. \tag{1}$$

This is signal separation, an inverse problem: one equation, two unknowns, unsolvable without assumptions. The standard one exploits the night: with no photosynthesis, $\mathrm{GPP} = 0$ and NEE is respiration alone, so a parametric or machine learning (ML) model fitted to nighttime data estimates $\mathrm{R_{eco}}$, which, when extrapolated through the day and subtracted from NEE, yields GPP (Reichstein et al., 2005b; Pastorello et al., 2020). This is a sensible first step (Figure 1). Unfortunately, it is not that simple. At night, the meteorological drivers vary far less than by day, so relationships are harder to learn and both covariate and concept shift occur across the night–day boundary, forcing the respiration model to extrapolate (Lasslop et al., 2010). The measurements are also worse: EC relies on turbulent mixing, often weak on calm, stably stratified nights, leaving the nighttime record noisier and more gap-ridden. Add abrupt regime shifts such as rain pulses and harvests (Nguyen

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2025), and the task calls for exactly what ML offers: flexibility with knowledge integration, robust extrapolation, transfer across space and time, and tolerance to gaps.

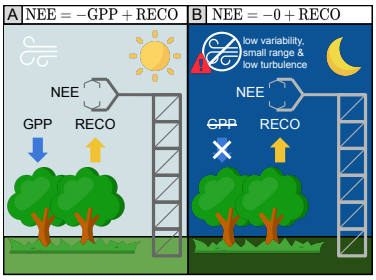


Figure 1: Flux partitioning over day and night. Only the net flux is measured, $\mathrm{NEE} = -\mathrm{GPP} + \mathrm{R_{eco}}$. The goal is to retrieve GPP and $\mathrm{R_{eco}}$. **(A) Day:** both components are active, and the split is undetermined. **(B) Night:** photosynthesis stops ($\mathrm{GPP} = 0$), so $\mathrm{NEE} = \mathrm{R_{eco}}$ — the constraint that makes partitioning tractable, yet nighttime is when the data is least reliable. The strongest constraint coincides with the weakest data.

Carbon fluxes have become an ML problem in their own right, with recent datasets and benchmarks for flux modeling, upscaling, and forecasting (Fortier et al., 2025; Nelson et al., 2024; Fries et al., 2026; Wang et al., 2025). Yet each rests on data that are themselves partitioning estimates—as targets or as calibration references—so none evaluates partitioning itself. The field's gold-standard tools are two parametric methods, nighttime (Reichstein et al., 2005a) and daytime (Lasslop et al., 2010) partitioning, integral to the FLUXNET data products. Their simple functional forms, fitted in moving windows, are known to be brittle across biomes and to miss phenomena such as rain-induced respiration pulses in drylands, which cover over a third of the global land surface (Nguyen et al., 2025). ML alternatives (Moffat et al., 2007; Desai et al., 2008; Tramontana et al., 2020; Trifunov et al., 2021; Zhan et al., 2022; Cohrs et al., 2024) drop the hand-chosen forms but carry their own risks—physical-law violations, poor out-of-distribution (OOD) generalization, and opacity (Reichstein et al., 2019; Marcus, 2018; Quionero-Candela et al., 2009; Sugiyama & Kawanabe, 2012; Rudin & Radin, 2019)—and none has entered the standardized pipelines, which would require characterizing their limitations and comparing them against existing methods across many sites and years. And that comparison is the elephant in the room. A developer (from the field or outside it) has no agreed metric, no common dataset, and, above all, no ground truth for the separated fluxes. Trying a new method alongside the established ones, to see what it fixes and what it breaks, means re-running each baseline on whatever data is at hand. Existing studies bear this out: they use shifting sets of sites and lean on synthetic data from a single mechanistic model of the authors' choosing without realistic noise or gaps (Moffat et al., 2007; Desai et al., 2008; Tramontana et al., 2020; Trifunov et al., 2021; Zhan et al., 2022; Cohrs et al., 2024). We address this with a dataset and evaluation protocol for carbon flux partitioning. Its core is synthetic ground truth from two structurally distinct land-surface models, QUINCY (Thum et al., 2019) and CABLE-POP (Haverd et al., 2018): because they output GPP and $\mathrm{R_{eco}}$, we can test directly whether a method recovers them, and using two independent lineages keeps a method from scoring well by emulating a single simulator while spanning a broad range of biomes. To keep the challenge realistic, we reintroduce measurement noise and gaps into the net ecosystem exchange (NEE) signal and impose abrupt rain-pulse regime shifts built from the flux-tower records. We call the benchmark and evaluation protocol **FluxPart-E**[1]. The main contributions of this work are:

- **A testbed for an unevaluable problem.** FluxPart-E supplies known GPP/$\mathrm{R_{eco}}$ ground truth from two structurally distinct land-surface models driven by 188 FLUXNET sites with measurement-realistic noise, gaps, and events. It is leakage-free by construction, since the target decomposition is synthetic and never exposed to any method.
- **An evaluation protocol, not just data.** We score the latent decomposition, not the NEE reconstruction, across six temporal scales and targeted stress regimes (data coverage, climate-zone edges, peak temperatures, rain pulses).
- **Ten methods, no winner.** Three parametric and seven ML methods: none dominates; rankings shift with flux, metric, and regime, and the two classes fail in complementary, identifiable ways.
- **A bridge to real data.** Inter-method disagreement patterns largely transfer to real EC data. Broad regimes carry over, rain pulses remain an identified gap. This is giving external validity and flagging where methods diverge most.
- **Open and extensible.** The benchmark dataset is released under CC-BY 4.0.

[1]The "E" denotes Evaluation. Read aloud, it is "Flux Party."

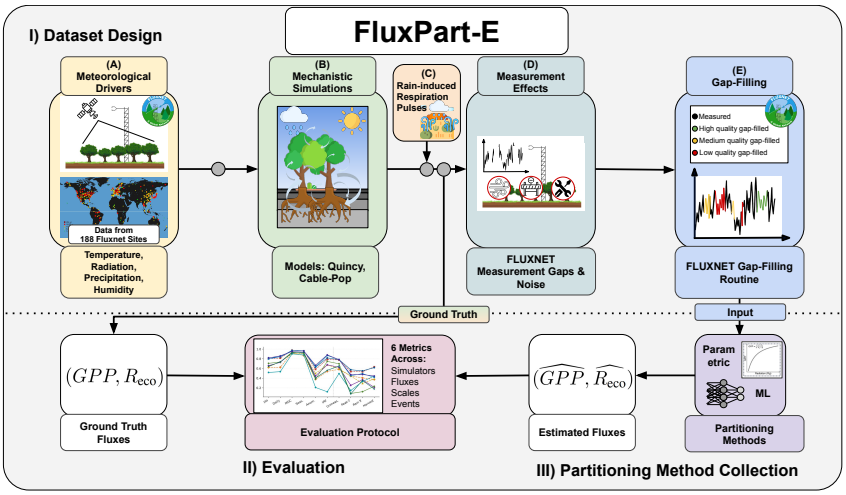


Figure 2: Overview of the FluxPart-E benchmark. **(I) Dataset design:** meteorological drivers from FLUXNET sites (A) drive two land-surface models (B) that output NEE with ground-truth $\mathrm{GPP}/\mathrm{R_{eco}}$; rain-induced respiration pulses (C) are added post hoc, then site-specific noise and gaps (D) and standard gap-filling (E) yield the partitioning inputs. **(II) Method collection:** parametric (NT, DT, TK) and ML (double machine learning (DML), neural network (NN)) methods (table 8). **(III) Evaluation:** on both simulators across temporal scales and event types (tables 5 and 6). Details in section 3.

2 Related Work

Machine-learning benchmarks for the environmental sciences have grown rapidly, both to spur new ML methods and to tackle problems with real-world impact, spanning sea-ice mapping (de Loë et al., 2025), seismic imaging (Deng et al., 2022), weather and climate emulation (Watson-Parris et al., 2022; Kaltenborn et al., 2023; Yu et al., 2025), land-surface prediction (Li et al., 2024), general physical dynamics (Ohana et al., 2024; Takamoto et al., 2023), and carbon-flux modelling (Fortier et al., 2025; Nelson et al., 2024). Many draw their ground truth from physical simulations rather than observations (Ohana et al., 2024; Takamoto et al., 2023; Watson-Parris et al., 2022; Kaltenborn et al., 2023; Yu et al., 2025), which supply complete, noise-free targets with full knowledge of the generating process—a reference observations cannot provide. FluxPart-E falls into this category. Most such benchmarks address *forward* problems, predicting an observable from inputs. Fewer pose *inverse* problems that recover latent quantities from observations, e.g. subsurface velocity from seismic waveforms (Deng et al., 2022), fields from PDE solutions (Takamoto et al., 2023), or sea-ice state from imagery (de Loë et al., 2025). Flux partitioning is inverse—GPP and $\mathrm{R_{eco}}$ are recovered from a single net flux—so FluxPart-E sits at the intersection: a simulation-grounded inverse-problem benchmark. It differs in one respect that is decisive here. In those other inverse benchmarks, the latent targets can in principle be measured (drilling a well, sampling an ice floe); the component fluxes cannot, so simulation is not a convenience but the only route to evaluation. This is why existing carbon-flux benchmarks remain forward-modeling efforts that in fact *consume* partitioning estimates (Fortier et al., 2025; Nelson et al., 2024): none has targeted the partitioning problem itself.

3 Methods

3.1 Dataset Design

Guiding Principle. FluxPart-E combines controlled conditions with realistic measurement characteristics. Ground truth for GPP and $\mathrm{R_{eco}}$ requires mechanistic land-surface models, while realism requires that their output resemble EC data. We therefore drive the models with meteorology from real flux-tower sites and reintroduce site-specific NEE noise, FLUXNET data gaps, and rain-induced respiration pulses. The result is a controlled setting with known ground truth that retains the variability, noise, and data limitations typical of FLUXNET observations.

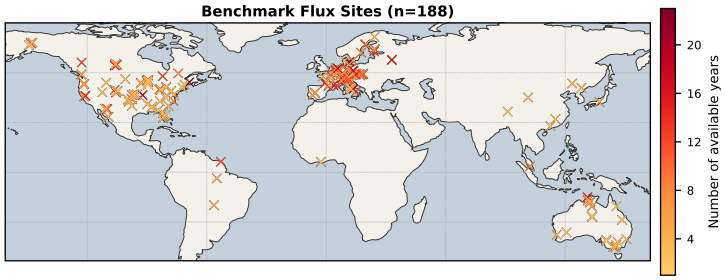


Figure 3: Geographic distribution of the $n = 188$ benchmark sites. Color indicates the number of available years of data per site. Coverage follows the well-documented geographic bias of the underlying FLUXNET network, which is a higher concentration of sites in North America and Europe and a sparser representation in Africa, South America, and Asia. (Meyer & Pebesma, 2022)

Overall Pipeline. The pipeline (fig. 2) runs in five steps (A–E), each detailed below.

(A) Meteorological Drivers. Each simulation is driven by the meteorology of a real flux-tower site, so the benchmark inherits realistic temporal variability and site-specific conditions rather than idealized forcings. The drivers are the key land-surface variables (air temperature, radiation, humidity, among others) taken from the processed FLUXNET forcing set of Nelson et al. (2024). We retain sites that have simulations from both QUINCY and CABLE-POP and enough directly measured NEE for evaluation, yielding $n = 188$ sites spanning 11 IGBP land-cover classes and 1403 site-years. The full selection criteria and site list are in Section A.1 (tables 2 and 3). The set covers a broad range of biomes and climates but inherits the well-documented geographic bias of FLUXNET toward North America and Europe, with sparser sampling in Africa, South America, and continental Asia (fig. 3) (Meyer & Pebesma, 2022). As FLUXNET coverage grows, we will update the dataset to better sample sparse regions.

(B) Mechanistic Simulations. The meteorological inputs drive two mechanistic land-surface models, QUINCY (Thum et al., 2019) and CABLE-POP (Haverd et al., 2018), each yielding NEE with ground-truth GPP and R_{eco}. Both simulate coupled carbon and nutrient cycles but come from independent lineages and differ structurally in vegetation dynamics and photosynthesis (Section A.2), so a method cannot win by emulating one simulator alone; evaluating against both spans a broader range of ecosystem dynamics.

(C) Event Design: Rain-induced Respiration Pulses. When dry soils are rewetted, microbial activity resumes abruptly and releases a short, intense burst of CO_2, known as the Birch effect. These pulses matter in water-limited ecosystems and are hard to capture: they are *non-steady-state*, depend on antecedent dryness and substrate build-up, and appear to a learned model as a short, high-magnitude departure from the otherwise smooth driver–flux relationship. Nguyen et al. (2025) compare parametric and ML approaches and find that all underestimate the pulses. We therefore include them as a targeted stress regime. We follow the formulation by Nguyen et al. (2025) and adapt it to operate on simulated fluxes and to integrate site-specific pulse annotations. In total, the dataset includes 18 sites with rain pulse events, accounting for 870 pulses. Full details are in Section A.3.

(D) Reintroduction of Measurement Effects and Data Gaps. To turn the simulator output into a realistic machine-learning target, we reimpose the measurement effects of a real eddy-covariance record. We add heavy-tailed Laplace noise to the simulated NEE, scaled by each site's FLUXNET random-uncertainty estimate and rescaled to the model's flux magnitude; the Laplace form matches the heavy-tailed error of EC measurements (Hollinger & Richardson, 2005). We then apply the site's FLUXNET quality mask to impose realistic data gaps. Together, these give each simulated site-year the noise structure and missing-data pattern of its real-world counterpart. Full definitions, variable names, and the coverage distribution are in Section A.4.

(E) Gap-Filling. Real partitioning is applied not to raw EC data but to gap-filled products: the standard FLUXNET pipeline reconstructs missing NEE before partitioning, and this step is itself a recognized source of additional uncertainty. To keep the benchmark realistic, we gap-fill the noisy simulated NEE with the same standard procedure: the marginal distribution sampling imple-

mented in the REddyProc library (Wutzler et al., 2018). This ensures consistency with established FLUXNET processing pipelines and defines the *unmeasured* evaluation regime on which we later assess robustness to gap-filling (table 6).

3.2 Evaluation Protocol

A partitioning method can be accurate in one setting and unreliable in another, so a single aggregate score reveals little. Our protocol therefore evaluates the recovered GPP and $\mathrm{R_{eco}}$ along three axes: across temporal scales, under targeted stress regimes, and with a complementary set of metrics. Together they localize *where* a method fails, not just *whether* it does.

Temporal Scales. Flux estimates are used from the half-hourly signal to year-to-year change, and aggregation matters: errors at one scale can cancel or accumulate at another. We evaluate at six scales: half-hourly (half-hourly (HH)), the base unit; daily (daily (D)); mean diurnal cycle (mean diurnal cycle (MDC)), for diurnal phase and amplitude; seasonal cycle (seasonal cycle (SC)), for phenology and carbon budgets; anomalies (anomalies (Anom)), departures from the seasonal course driven by events such as heatwaves and droughts; and inter-annual variability (inter-annual variability (IAV)), the notoriously hardest scale (Jung et al., 2019). Precise definitions are in Appendix table 5.

Event Types. Beyond aggregate scales, we evaluate three regimes known to stress partitioning. *Rain-induced respiration pulses* are scored on a window around each event, testing whether a method adapts to the abrupt change and whether it does so too early or too late. *Unmeasured* timesteps are those that were gap-filled rather than directly observed, testing robustness to missing or corrupted data. *Peak-temperature* timesteps are those whose air temperature exceeds the site-year's maximum nighttime temperature, testing methods exactly where a night-calibrated respiration model must extrapolate beyond its fitted range. Appendix table 6 summarizes the regimes.

Metrics. No single number captures how a partition can go wrong, so we report metrics that isolate distinct aspects: *bias* for systematic over- or under-estimation (level), squared Pearson correlation (r^2) for temporal pattern, and *cRMSE* for the remaining error in the dynamics once the mean offset is removed. *RMSE* is their composite ($\mathrm{RMSE}^2 = \mathrm{bias}^2 + \mathrm{cRMSE}^2$), reported as the familiar total-error summary, and *NSE* normalizes this total by the reference variance to give a scale-comparable skill score [2]. Because flux magnitudes differ between models and real sites, bias, RMSE, and cRMSE are informative only within a model, whereas r^2 and NSE are scale-robust across them. Full definitions are in Appendix table 7.

3.3 Partitioning Methods

We benchmark ten methods chosen to cover the main partitioning families. We provide three parametric baselines and seven ML methods that vary in structure, loss, features, knowledge integration, and transfer. A full specification is given in Appendix table 8.

Parametric Baselines. The field's long-standing standard tools fit simple functional forms through moving windows over the data. The nighttime method (NT (Reichstein et al., 2005b)) extrapolates a nighttime respiration function into the day, and the daytime method (DT (Lasslop et al., 2010)) includes a light-response function for GPP. We additionally include a more recent modified daytime variant (TK (Keenan et al., 2019)). These assumption-driven baselines serve as the reference against which we judge new methods.

Machine Learning Baselines. We include a selection of machine learning methods suggested in the literature: a method based on DML, two different NN architectures, and several extensions. $\mathrm{NN_{night}}$ (the no-sif architecture from Zhan et al. (2022)) imposes minimal structure: no explicit form for GPP, only the constraint that it vanishes at night. The remaining methods use the light-use-efficiency factorization (Tramontana et al., 2020). Within that family, we vary one ingredient at a time: the loss (NN_{L1}, an L_1 variant matched to the Laplacian noise model and our reference architecture, versus the MSE baseline $\mathrm{NN_{base}}$). $\mathrm{NN_{FE}}$ uses various engineered features: light-use-efficiency proxies, nighttime dynamics, and soil-moisture memory. $\mathrm{NN_{KG}}$ leverages knowledge-guided regularization constraining $\mathrm{R_{eco}}$ with the parametric NT product. $\mathrm{NN_{PT}}$ uses transfer over time by multi-year

[2] We note that Nash-Sutcliffe efficiency (NSE) is identical to the coefficient of determination R^2 score but less commonly used in the domain. To avoid confusion with the squared Pearson correlation r^2, we keep NSE.

pretraining before per-site-year fine-tuning. Finally, DML (Cohrs et al., 2024) is a causal, double-machine-learning approach that estimates GPP through a partially linear model with gradient-boosted nuisance functions. Architectures and training details are in Section B.

3.4 SHARP BITS

FluxPart-E targets the retrieval of latent fluxes rather than ordinary supervised prediction, so several habits of standard ML setups do not apply cleanly. We collect the most consequential here. They share one resolution: the benchmark fixes only the evaluation: every method is scored on the recovered GPP/R_{eco} over the full site-year against known ground truth. This leaves the method-internal choices open, because in a retrieval problem those choices are part of the method.

Using the Benchmark: Splits and Hold-out Because a method consumes the full site-year NEE to infer the decomposition, there is no held-out split in the supervised sense; internal splitting (for model selection or early stopping) and any cross-site or cross-year hold-out are method-dependent and left to the user (we discuss this more thoroughly in Section A.8).

Circularity through Knowledge Integration. Since the ground truth comes from known mechanistic models, why not just run the model to recover the partition? This is no shortcut: the models are complex with long spin-ups, and calibrating them to the same noisy, gap-ridden NEE is itself an unsolved inverse problem—a legitimate contribution we explicitly invite. Integrating domain knowledge or engineering features is encouraged, not penalized: a method may exploit mechanistic structure, governing equations, or other partitioning products as regularization targets, as one of our knowledge-guided variants does.

The Golden Rule. The benchmark enforces exactly one rule: *no method may have access to the latent ground truths of* GPP *and* R_{eco}. Everything else, including physical priors, parametric products, cross-site transfer, is fair game.

4 EXPERIMENTS AND RESULTS

We evaluate all ten methods on both simulators across the full protocol (section 3.2): metrics, scales, and events. We first show that no method suffices—skill collapses where it matters, and none dominates (section 4.1)—then localize the failures by data coverage and stress regime, and finally show that method-disagreement patterns largely transfer to real data, giving the benchmark external validity.[3]

4.1 NO OFF-THE-SHELF METHOD SUFFICES

At half-hourly resolution, the finest scale, we ask whether any method dominates across site-years—a site-year (one calendar year at one site) is our unit of evaluation. Table 9 reports the share of site-years for which each method ranks first on NSE (dynamics) and bias (structural over-/under-estimation). No method dominates. Leadership flips between simulators: NN_{KG} leads the component fluxes on QUINCY (51% best-share for GPP, 54% for R_{eco} on NSE) but falls to 11% for GPP on CABLE-POP, where NN_{PT} leads instead (54%). And the dynamics leaders are not the bias leaders—ML methods lead on NSE, yet the parametric NT keeps the bias edge (up to 25% best-share, R_{eco} on QUINCY). The share still hides the *magnitude* of these gaps, so we turn to the median squared Pearson correlation (r^2) between estimated and ground-truth fluxes across scales and events (fig. 4). Overall, the two simulators agree, with challenging scales and conditions more pronounced in CABLE-POP. Correlation is high at the half-hourly, daily, MDC, and SC scales (unmeasured points slightly lower), but plunges at the scales that matter most—anomalies, IAV, peak temperatures, and rain pulses. The NN_{L1}-family extensions (NN_{KG}, NN_{PT}) perform best across scales and events and track each other closely; the remaining ML methods are mixed, with the minimum-structure NN_{night}, DML, and NN_{FE} lowest (NN_{night} lacks structural constraints; DML and NN_{FE} rely on engineered, partly gap-filled features). Parametric methods fall between, DT mostly below NT. ML clearly leads on

[3]We provide an interactive dashboard for local deployment to these results with re-rankable method tables and distribution views across metrics, temporal scales, and stress regimes. It can be built from the code and data in `https://github.com/AnonymousAuthor1194/anonymous-submission-fluxpartbench-results-dashboard/`

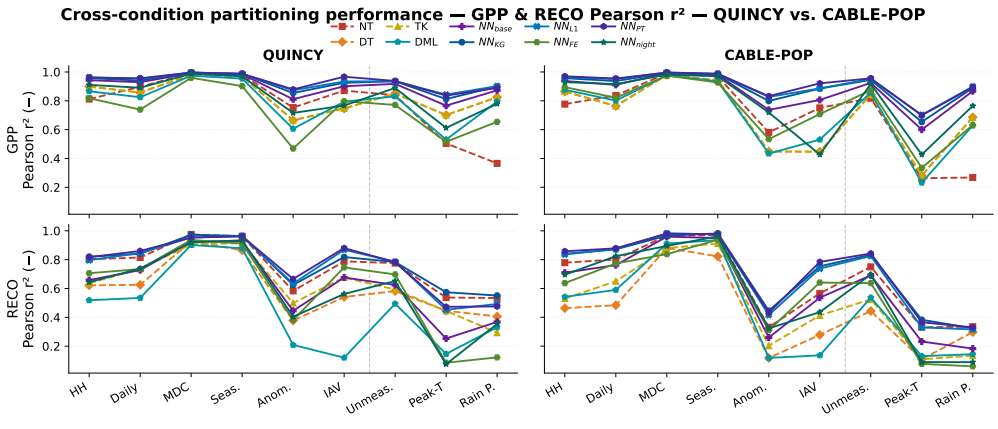


Figure 4: Cross-condition partitioning performance. Median squared Pearson correlation (r^2) between estimated and ground-truth GPP (top) and RECO (bottom) for QUINCY (left) and CABLE-POP (right), across conditions: half-hourly (HH), daily, mean diurnal cycle (MDC), seasonal cycle (Seas.), anomalies (Anom.), inter-annual variability (IAV), unmeasured/gap-filled (Unmeas.), peak-temperature (Peak-T), and rain pulses (Rain P.). The dotted line separates scale-based (left) from stress-event (right) conditions.

GPP; on $\mathrm{R_{eco}}$, NT is often on par. Under the stricter NSE, the collapse is starker: medians fall below zero (worse than the site-year mean) on anomalies, IAV, peak-temperature, and pulses (Section C.2 with per-method distributions). Together, this is the benchmark's central negative result: parametric and ML methods alike share a gap on exactly the regimes that matter, and none closes it. The rest of this section localizes where and why.

4.2 Stress-regime analysis: identifying failure modes

We now examine the known pain points in more detail, restricting each comparison to a relevant subset of methods. We use NT and DT as parametric references and NN_{L1} as the reference LUE-based network (we omit $\mathrm{NN_{base}}$, which NN_{L1} dominates, differing only by an MSE rather than L_1 loss).

Coverage does not matter equally across IGBPs. ML models are data-hungry (LeCun et al., 2015; Kaplan et al., 2020), yet towers give only a few, often gap-ridden years per site (fig. 9), and gap-filling adds its own uncertainty (Moffat et al., 2007; Vekuri et al., 2023). We bin site-years up to 60% coverage to isolate the limited-data regime and compare across IGBP classes (fig. 5; remaining classes in figs. 16 to 18), showing $\mathrm{NN_{PT}}$ and $\mathrm{NN_{KG}}$ (which target this regime) alongside the structurally distinct DML and $\mathrm{NN_{night}}$. Parametric and ML methods split clearly. NT and DT perform worse overall but stay stable across coverage bins. Their structural bias recovers the diurnal cycle regardless of data amount, whereas ML methods degrade in the lowest bins, most in crop and grassland and more for $\mathrm{R_{eco}}$ than GPP (barely at all for evergreen needleleaf forest). Pretraining does not help: $\mathrm{NN_{PT}}$ shows no gain over NN_{L1}. On EBF, SAV, and OSH, all methods struggle regardless of coverage. This is a data-independent failure that points to climate rather than sample size, which we examine next.

Climate Zones. These struggles point to warm zones, consistent with reports that partitioning fails under drought and high-VPD extremes (Zhan et al., 2022; Kira et al., 2021; Jung et al., 2024). Mapping performance across climate space for NT, DT, and NN_{L1} (Appendix fig. 19; both simulators comparable), $\mathrm{R_{eco}}$ degrades toward warm zones for all three, most strongly for DT, while GPP degrades mainly for NT and NN_{L1} largely solves it except at the warm-dry and very-wet-warm margins. The methods have distinct failure geographies, which the benchmark makes visible.

Peak temperatures. ML models struggle to extrapolate beyond their training range (Haley & Soloway, 1992): if $\mathrm{R_{eco}}$ is constrained mainly by nighttime data and GPP compensates by day, methods should fail at daytime temperatures above the site-year's highest nighttime temperature (Kira et al., 2021; Zhan et al., 2022). We split days by that threshold and compare GPP/$\mathrm{R_{eco}}$ error, adding $\mathrm{NN_{KG}}$ (whose NT anchor targets this regime) to NN_{L1}. NEE stays unbiased in both groups

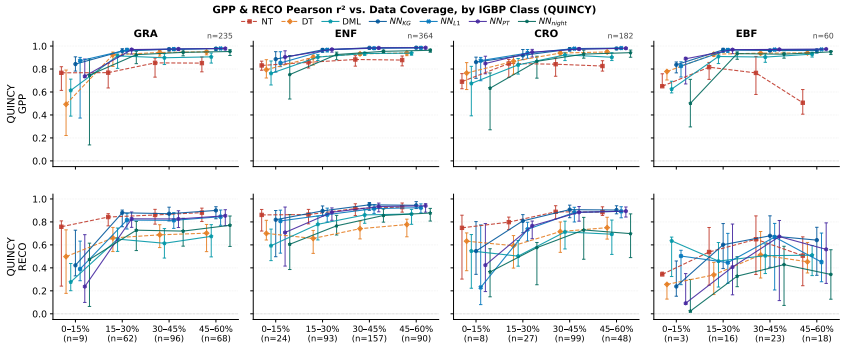


Figure 5: Partitioning performance by data coverage and IGBP class (QUINCY). Half-hourly r^2 for a method subset (rationale in section C.3), by flux and coverage bin (up to 60%) for grassland (GRA), evergreen needleleaf forest (ENF), cropland (CRO), and evergreen broadleaf forest (EBF). Error bars show IQR; site-years n per bin in parentheses, total per class at top. Interpret bins with $n < 10$ with caution.

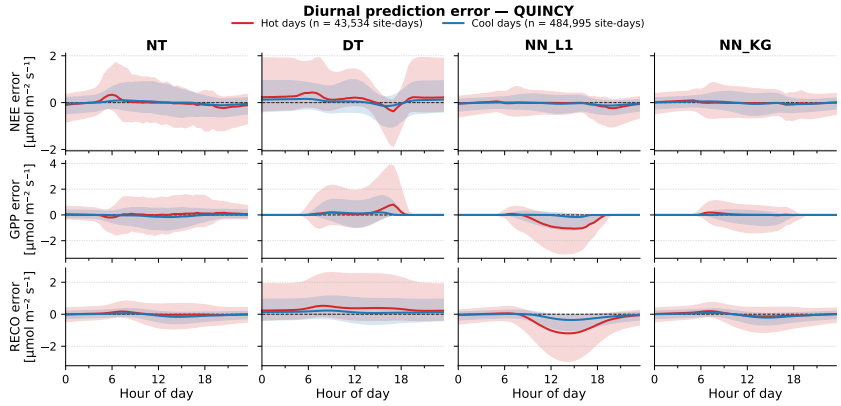


Figure 6: Diurnal prediction error (QUINCY) for NT, DT, NN_{L1}, and $\mathrm{NN}_{\mathrm{KG}}$, split into warm and cool days (warm: daytime temperature above the site-year's maximum nighttime temperature). Lines show median error per hour-of-day bin across sites; bands show IQR.

and simulators—the failure is in the partition. On QUINCY (fig. 6), NT stays unbiased on warm days while NN_{L1} underestimates both fluxes (compensating in NEE), confirming the extrapolation hypothesis; on CABLE-POP the pattern reverses (fig. 21), with NT overestimating and NN_{L1} mostly unbiased. $\mathrm{NN}_{\mathrm{KG}}$ inherits its NT anchor, so it helps on QUINCY but harms on CABLE-POP—the bias direction is thus simulator-dependent, and future methods need an adaptive NT anchor.

Rain pulses. Nguyen et al. (2025) report that both standard and ML methods fail to capture rain-induced respiration pulses; we test whether the benchmark reproduces this. Over a five-day window around pulse onset (fig. 7; CABLE-POP in fig. 22), DT most strongly underestimates post-pulse $\mathrm{R}_{\mathrm{eco}}$ on both simulators, and NT underestimates post-pulse GPP on QUINCY. NN_{L1} has the lowest absolute error, but no method (ML included) captures the pulses well. Single events explain why (Appendix fig. 23): NT and DT adapt slowly through the pulse because of their window fitting, while NN_{L1} and $\mathrm{NN}_{\mathrm{KG}}$ jump at day boundaries, reflecting their daily flux-proxy inputs (Tramontana et al., 2020). $\mathrm{NN}_{\mathrm{FE}}$, using a moving-window feature, can capture the rise smoothly, but not robustly, since at other events all methods miss the pulse, so it is not yet a solution.

4.3 Transfer to real data.

Finally, we ask whether the benchmark's findings carry meaning for real flux-tower data. Direct verification is impossible as the absence of ground-truth GPP and $\mathrm{R}_{\mathrm{eco}}$ at real towers is exactly

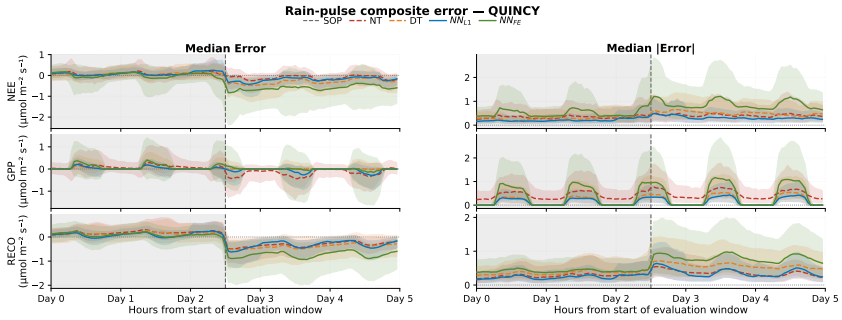


Figure 7: Pulse-event composite error (QUINCY) for NT, DT, NN_{L1}, and $\mathrm{NN}_{\mathrm{FE}}$: median error (left) and median absolute error (right) for NEE, GPP, and RECO over a five-day window around pulse onset (dashed line, Day 2.5; shaded region, pre-pulse baseline). Bands show IQR across events.

Table 1: The four stress regimes the benchmark isolates map to distinct, well-posed machine-learning problems. In each, a natural first attempt was already among our baselines and proved insufficient; the listed directions are candidates, not prescriptions.

Stress regime	**ML problem**	**Directions to explore**
Sparse data (low coverage)	Few-shot generalization; learning under missingness	Meta-learning across site-years; data-efficient inductive biases; learning gap-filling jointly rather than as a fixed upstream step (e.g. physics-constrained generative imputation).
Climate zones	Spatial transfer under low local variability	Invariant or causal representations; spatial transfer; robustness in under-sampled climate corners.
Temperature extremes	Reliable extrapolation beyond the training support	Physics-guided constraints; OOD-aware uncertainty.
Rain pulses	Learning rapid dynamics without oversmoothing	Sub-daily sequence models with memory; event-aware losses that do not regress rare events to the mean; rapid adaptation.

what motivated a simulation-based benchmark. One quantity is computable on both, however: the disagreement between methods, which we expect to grow in the challenging regimes identified above and to follow similar patterns on simulated and measured data. It largely does. The broad regime structure transfers, the warm-zone failure cluster in climate space and the peak-temperature disagreement structure reproduce on real data, and coverage behaves consistently down into the sparse regime, where the benchmark partly under- or overestimates the real spread. The clear exception is event-specific dynamics: transfer breaks down for rain pulses, where both simulators smooth transients that real data show—a limitation to close in future iterations. Full analyses in Section C.4.

5 Discussion

Flux partitioning is a real, high-stakes ML problem that has been hard even to pose: two latent fluxes recovered from a single observed constraint, under distribution shift and sparse, noisy supervision, with no ground truth on the data where methods are deployed. GPP and $\mathrm{R}_{\mathrm{eco}}$ products underpin carbon-cycle science, model benchmarking, and global upscaling, yet the problem has been effectively unevaluable. Our contribution is not any single result but a framework that makes it measurable—shared synthetic ground truth, a multi-scale evaluation protocol, a set of baselines and extensions, and evidence that method-disagreement patterns largely transfer from simulated to real EC data—letting the benchmark locate where partitioning is unreliable, which real data alone cannot. Each stress regime it isolates maps to a distinct, well-posed ML problem for which a natural first attempt was already among our baselines and proved insufficient. Table 1 summarizes these open problems and candidate directions.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will release the benchmark dataset (meteorological drivers, synthetic $\mathrm{GPP}/\mathrm{R}_{\mathrm{eco}}$ ground truth, measurement masks, and injected noise), the evaluation protocol, reference implementations of all ten methods, and the code needed to reproduce every result in this paper. For review, an anonymized version of the data is already available at `https://huggingface.co/datasets/anonymous-author-fluxpartbench/anonymous-submission-fluxpartbench-data`; the supplementary material includes code for running and evaluating a baseline partitioning method; and an interactive results dashboard with all results of the submission, deployable locally, can be built from the code and data at `https://github.com/AnonymousAuthor1194/anonymous-submission-fluxpartbench-results-dashboard/`. The dataset will receive a persistent DOI on acceptance. Data are released under CC-BY 4.0 and code under the MIT License, with attribution to the FLUXNET/ICOS towers and the underlying forcing dataset.

ETHICS STATEMENT

FluxPart-E inherits the geographic sampling of the FLUXNET/ICOS network that supplies its meteorological drivers, and requiring output from both land-surface models together with quality-screened measured NEE narrows it further. The 188 sites are concentrated in North America and Europe, which together account for 84% of sites and 90% of site-years; only 8% lie in the tropics, with a single site in the Amazon basin and none in the Congo (fig. 3) (Meyer & Pebesma, 2022). Two of the largest terrestrial carbon reservoirs are thus effectively absent, and results should not be assumed to transfer to humid-tropical forests or African ecosystems. This under-representation falls disproportionately on the tropics and much of the Global South—regions central to the global carbon cycle yet least represented in the observational networks the field depends on. We therefore report performance stratified by climate zone and IGBP class rather than as a single global score. The benchmark does cover semi-arid and dryland systems (savannas, shrublands, and the rain-pulse regime), so this gap is specifically one of humid-tropical and African coverage rather than of climate extremes in general.

Because partitioned GPP and $\mathrm{R}_{\mathrm{eco}}$ feed carbon-cycle science, model benchmarking, and global upscaling, methods or products tuned on this benchmark could, if applied uncritically, systematically misrepresent the very regions least able to supply corrective local data. We therefore regard the coverage gap as a limitation to be closed rather than a fixed property of the benchmark, and treat improved sampling of these under-served regions as a priority for future iterations as FLUXNET/ICOS coverage grows. The dataset is released under CC-BY 4.0 and the code under the MIT License, keeping the benchmark broadly accessible and open to such contributions.

AI USE STATEMENT

We used generative AI tools to assist with LaTeX formatting for the conference template and with copy-editing for readability. We have not used generative AI tools to generate synthetic datasets, develop theoretical models, formulate mathematical claims, implement methods, or interpret results. We have reviewed all AI-assisted text. We take responsibility for the final content of this work, including text, claims, and artifacts.

REFERENCES

Brian Amiro. Fluxnet2015 CA-SF2 Saskatchewan - Western Boreal, forest burned in 1989, a. URL `https://doi.org/gr2s39`.

Brian Amiro. Ameriflux FLUXNET-1F CA-SF1 Saskatchewan - Western Boreal, forest burned in 1977, b. URL `https://doi.org/hchx7t`.

Brian Amiro. Fluxnet2015 CA-SF3 Saskatchewan - Western Boreal, forest burned in 1998, c. URL `https://doi.org/gr2s4b`.

Christoph Ammann. Fluxnet2015 CH-Oe1 Oensingen grassland. URL `https://doi.org/gr2s6s`.

M. Arain. Ameriflux FLUXNET-1F CA-TPD Ontario - Turkey Point Mature Deciduous, a. URL https://doi.org/gr2s2c.

M. Altaf Arain. Fluxnet2015 CA-TP4 Ontario - Turkey Point 1939 Plantation White Pine, b. URL https://doi.org/gr2s4g.

Stefan Arndt, Nina Hinko-Najera, and Anne Griebel. Fluxnet2015 AU-Wom Wombat. URL https://doi.org/gr2s8k.

Svante Arrhenius. Über die reaktionsgeschwindigkeit bei der inversion von rohrzucker durch säuren. *Zeitschrift für physikalische Chemie*, 4(1):226–248, 1889.

Mika Aurela, Annalea Lohila, Juha-Pekka Tuovinen, Juha Hatakka, Juuso Rainne, Timo Mäkelä, and Tuomas Lauria. Fluxnet2015 FI-Lom Lompolojankka. URL https://doi.org/gr2s85.

John Baker and Tim Griffis. Ameriflux FLUXNET-1F US-Ro4 Rosemount Prairie, a. URL https://doi.org/gr2s28.

John Baker and Tim Griffis. Ameriflux FLUXNET-1F US-Ro6 Rosemount I18_North, b. URL https://doi.org/gr2s29.

John Baker and Tim Griffis. Ameriflux FLUXNET-1F US-Ro5 Rosemount I18_South, c. URL https://doi.org/gr2sxs.

John Baker, Tim Griffis, and Timothy Griffis. Ameriflux FLUXNET-1F US-Ro1 Rosemount- G21. URL https://doi.org/gr2s27.

D. Baldocchi, E. Falge, L. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, R. Evans, J. Fuentes, A. Goldstein, G. Katul, B. Law, X. Lee, Y. Malhi, T. Meyers, W. Munger, W. Oechel, Paw U,K.T., K. Pilegaard, H.P. Schmid, R. Valentini, S. Verma, T. Vesala, K. Wilson, and S. 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. ISSN 0003-0007. doi: 10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2.

Dennis D. Baldocchi. How eddy covariance flux measurements have contributed to our understanding of global change biology. *Global Change Biology*, 26(1):242–260, 2020. doi: https://doi.org/10.1111/gcb.14807. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.14807.

Paul Berbigier and Denis Loustau. Fluxnet2015 FR-LBr Le Bray. URL https://doi.org/gr2s7d.

Jason Beringer and Lindsay Hutley. Fluxnet2015 AU-Stp Sturt Plains, a. URL https://doi.org/hchx7q.

Jason Beringer and Lindsay Hutley. Fluxnet2015 AU-How Howard Springs, b. URL https://doi.org/hchx7m.

Jason Beringer and Lindsay Hutley. Fluxnet2015 AU-DaP Daly River Savanna, c. URL https://doi.org/gr2s6n.

Jason Beringer and Lindsay Hutley. Fluxnet2015 AU-RDF Red Dirt Melon Farm, Northern Territory, d. URL https://doi.org/gr2s8d.

Jason Beringer and Lindsay Hutley. Fluxnet2015 AU-Dry Dry River, e. URL https://doi.org/gr2s79.

Jason Beringer and Prof. Lindsay Hutley. Fluxnet2015 AU-DaS Daly River Cleared, f. URL https://doi.org/gr2s6m.

Jason Beringer and Jeffery Walker. Fluxnet2015 AU-Ync Jaxa. URL https://doi.org/gr2s8m.

Jason Beringer, Shaun Cunningham, Patrick Baker, Timothy Cavagnaro, Ralph MacNally, Ross Thompson, and Ian McHugh. Fluxnet2015 AU-Rig Riggs Creek, a. URL https://doi.org/hchx7p.

Jason Beringer, Lindsay Hutley, David McGuire, and Paw U. Fluxnet2015 AU-Wac Wallaby Creek, b. URL https://doi.org/gr2s6r.

Dave Billesbach, James Bradford, and Margaret Torn. Fluxnet2015 US-AR2 ARM USDA UNL OSU Woodward Switchgrass 2, a. URL https://doi.org/gr2s55.

Dave Billesbach, James Bradford, and Margaret Torn. Fluxnet2015 US-AR1 ARM USDA UNL OSU Woodward Switchgrass 1, b. URL https://doi.org/gr2s54.

Dave Billesbach, Lara Kueppers, Margaret Torn, and Sebastien Biraud. Ameriflux FLUXNET-1F US-A32 ARM-SGP Medford hay pasture, c. URL https://doi.org/gr2s2d.

Sebastien Biraud, Marc Fischer, Stephen Chan, and Margaret Torn. Ameriflux FLUXNET-1F US-ARM ARM Southern Great Plains site- Lamont. URL https://doi.org/gr2szg.

T. Andrew Black. Fluxnet2015 CA-Obs Saskatchewan - Western Boreal, Mature Black Spruce, a. URL https://doi.org/gr2s35.

T. Andrew Black. Fluxnet2015 CA-Oas Saskatchewan - Western Boreal, Mature Aspen, b. URL https://doi.org/gr2s34.

Thomas Black. Ameriflux FLUXNET-1F CA-LP1 British Columbia - Mountain pine beetle-attacked lodgepole pine stand, c. URL https://doi.org/gr2sxv.

Peter Blanken, Russel Monson, Sean Burns, David Bowling, Andrew Turnipseed, and John Knowles. Ameriflux FLUXNET-1F US-NR1 Niwot Ridge Forest (LTER NWT1). URL https://doi.org/gr2szw.

Rosvel Bracho and Charles Hinkle. Ameriflux FLUXNET-1F US-KS3 Kennedy Space Center (salt marsh). URL https://doi.org/gr2s25.

Nathaniel Brunsell. Ameriflux FLUXNET-1F US-KFS Kansas Field Station, a. URL https://doi.org/gr2s24.

Nathaniel Brunsell. Ameriflux FLUXNET-1F US-KLS Kansas Land Institute, b. URL https://doi.org/gr2szh.

Jean-Christophe Calvet, William Maurel, and Alexandre Paci. Etc l2 fluxnet (half-hourly) from toulouse, 2017-12-31–2021-12-31, 2022. URL https://hdl.handle.net/11676/8tfWuUDkx_e4DUB1Njld66C5.

Alessandro Cescatti, Barbara Marcolla, Roberto Zorer, and Damiano Gianelle. Fluxnet2015 IT-La2 Lavarone2. URL https://doi.org/gr2s9d.

Jiquan Chen and Housen Chu. Fluxnet2015 US-WPT Winous Point North Marsh, a. URL https://doi.org/gr2s6f.

Jiquan Chen and Housen Chu. Fluxnet2015 US-CRT Curtice Walter-Berger cropland, b. URL https://doi.org/gr2s6g.

Jiquan Chen, Housen Chu, and Asko Noormets. Fluxnet2015 US-Oho Oak Openings. URL https://doi.org/gr2s5n.

Andreas Christen and Sara Knox. Ameriflux FLUXNET-1F CA-DBB Delta Burns Bog. URL https://doi.org/gr2sz9.

James Cleverly and Derek Eamus. Fluxnet2015 AU-TTE Ti Tree East, a. URL https://doi.org/gr2s8h.

James Cleverly and Derek Eamus. Fluxnet2015 AU-ASM Alice Springs, b. URL https://doi.org/gr2s76.

Kai-Hendrik Cohrs, Gherardo Varando, Nuno Carvalhais, Markus Reichstein, and Gustau Camps-Valls. Causal hybrid modeling with double machine learning—applications in carbon flux modeling. *Machine Learning: Science and Technology*, 5(3):035021, jul 2024. doi: 10.1088/2632-2153/ad5a60. URL https://doi.org/10.1088/2632-2153/ad5a60.

Lily de Loë, David A. Clausi, and K. Andrea Scott. Fusing ice surface temperature with the ai4arctic dataset for improved deep learning-based sea ice mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:24766–24782, 2025. doi: 10.1109/JSTARS.2025.3610260.

Chengyuan Deng, Shihang Feng, Hanchen Wang, Xitong Zhang, Peng Jin, Yinan Feng, Qili Zeng, Yinpeng Chen, and Youzuo Lin. Openfwi: Large-scale multi-structural benchmark datasets for full waveform inversion. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 6007–6020. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/27d3ef263c7cb8d542c4f9815a49b69b-Paper-Datasets_and_Benchmarks.pdf.

Ankur Desai. Ameriflux FLUXNET-1F US-CS1 Central Sands Irrigated Agricultural Field, a. URL https://doi.org/gr2s2q.

Ankur Desai. Ameriflux FLUXNET-1F US-CS4 Central Sands Irrigated Agricultural Field, b. URL https://doi.org/gr2s2t.

Ankur Desai. Fluxnet2015 US-WCr Willow Creek, c. URL https://doi.org/gr2s5w.

Ankur Desai. Ameriflux FLUXNET-1F US-CS3 Central Sands Irrigated Agricultural Field, d. URL https://doi.org/gr2s2s.

Ankur Desai. Fluxnet2015 US-Syv Sylvania Wilderness Area, e. URL https://doi.org/gr2s5r.

Ankur Desai. Fluxnet2015 US-Los Lost Creek, f. URL https://doi.org/gr2s49.

Ankur Rashmikant Desai, Andrew D. Richardson, Antje Maria Moffat, Jens Kattge, D. Hollinger, Alan G. Barr, Eva Falge, Asko Noormets, Dario Papale, Markus Reichstein, and Vanessa J. Stauch. Cross-site evaluation of eddy covariance GPP and RE decomposition techniques. *Agricultural and Forest Meteorology*, 148:821–838, 2008.

Han Dolman, Dimmie Hendriks, Frans-Jan Parmentier, Luca Belelli Marchesini, Joshua Dean, and Ko Van Huissteden. Fluxnet2015 NL-Hor Horstermeer. URL https://doi.org/gr2s7r.

Gang Dong. Fluxnet2015 CN-Cng Changling. URL https://doi.org/gr2s8n.

Eugenie Euskirchen. Ameriflux FLUXNET-1F US-BZB Bonanza Creek Thermokarst Bog, a. URL https://doi.org/gr2s2f.

Eugenie Euskirchen. Ameriflux FLUXNET-1F US-BZF Bonanza Creek Rich Fen, b. URL https://doi.org/gr2s2g.

Cacilia Ewenz, Robert Stevens, and Gary Grigson. Fluxnet2015 AU-Lox Loxton. URL https://doi.org/hchx7n.

Christophe Flechard, Yannick Fauvel, and Yannick Hamon. Etc l2 fluxnet (half-hourly) from mejusseaume, 2018-12-31–2021-12-31, 2022. URL https://hdl.handle.net/11676/yyNe-l1tZUuf3DNCFqdkdkmf.

Matthew Fortier, Mats Leon Richter, Oliver Sonnentag, and Christopher Pal. Carbonsense: A multimodal dataset and baseline for carbon flux modelling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l8zRnvD95l.

John Frank, George Valentine, and Bill Massman. Ameriflux FLUXNET-1F US-GLE GLEES. URL https://doi.org/gr2szs.

Thomas Friborg, Rasmus Jensen, and Lars Rasmussen. Etc l2 fluxnet (half-hourly) from voulundgaard, 2019-12-31–2022-12-31, 2023. URL https://hdl.handle.net/11676/06IbYERibemtvD4qVHZXMQ2p.

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

Anya Fries, Jacob A Nelson, Martin Jung, Markus Reichstein, and Jonas Peters. Fluxtrapolation: A benchmark on extrapolating ecosystem fluxes, 2026. URL https://arxiv.org/abs/2605.19812.

Sarah Goslee. Ameriflux FLUXNET-1F US-HWB USDA ARS Pasture Sytems and Watershed Management Research Unit- Hawbecker Site. URL https://doi.org/gr2s2z.

Mike Goulden. Fluxnet2015 BR-Sa3 Santarem-Km83-Logged Forest. URL https://doi.org/gr2s3n.

P.J. Haley and D. Soloway. Extrapolation limitations of multilayer feedforward neural networks. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, volume 4, pp. 25–30 vol.4, 1992. doi: 10.1109/IJCNN.1992.227294.

Chad Hanson, Chris Still, and Bev Law. Ameriflux FLUXNET-1F US-Me2 Metolius mature ponderosa pine. URL https://doi.org/gr2szj.

V. Haverd, B. Smith, L. Nieradzik, P. R. Briggs, W. Woodgate, C. M. Trudinger, J. G. Canadell, and M. Cuntz. A new version of the cable land surface model (subversion revision r4601) incorporating land use and land cover change, woody vegetation demography, and a novel optimisation-based approach to plant coordination of photosynthesis. *Geoscientific Model Development*, 11(7):2995–3026, 2018. doi: 10.5194/gmd-11-2995-2018. URL https://gmd.copernicus.org/articles/11/2995/2018/.

D. Y. Hollinger and A. D. Richardson. Uncertainty in eddy covariance measurements and its application to physiological models. *Tree Physiology*, 25(7):873–885, 07 2005. ISSN 0829-318X. doi: 10.1093/treephys/25.7.873. URL https://doi.org/10.1093/treephys/25.7.873.

Adrien Jacotot, Jean-Baptiste Paroissien, and Laurent Perdereau. Etc l2 fluxnet (half-hourly) from la guette, 2016-12-31–2021-12-31, 2022. URL https://hdl.handle.net/11676/vaj-43TmKB8-mCvwFKUmlGUD.

M. Jung, J. Nelson, M. Migliavacca, T. El-Madany, D. Papale, M. Reichstein, S. Walther, and T. Wutzler. Technical note: Flagging inconsistencies in flux tower data. *Biogeosciences*, 21(7):1827–1846, 2024. doi: 10.5194/bg-21-1827-2024. URL https://bg.copernicus.org/articles/21/1827/2024/.

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, May 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0076-8. URL https://doi.org/10.1038/s41597-019-0076-8.

Julia Kaltenborn, Charlotte Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. Climateset: A large-scale climate model dataset for machine learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 21757–21792. Curran Associates, Inc., 2023. doi: 10.52202/075280-0952. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/44a6769fe6c695f8dfb347c649f7c9f0-Paper-Datasets_and_Benchmarks.pdf.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Trevor F. Keenan, Mirco Migliavacca, Dario Papale, Dennis Baldocchi, Markus Reichstein, Margaret Torn, and Thomas Wutzler. Widespread inhibition of daytime ecosystem respiration. *Nature Ecology & Evolution*, 3(3):407–415, Mar 2019. ISSN 2397-334X. doi: 10.1038/s41559-019-0809-2. URL https://doi.org/10.1038/s41559-019-0809-2.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2017. URL https://api.semanticscholar.org/CorpusID:4800342.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

O. Kira, C. Y-Y. Chang, L. Gu, J. Wen, Z. Hong, and Y. Sun. Partitioning net ecosystem exchange (nee) of co2 using solar-induced chlorophyll fluorescence (sif). *Geophysical Research Letters*, 48(4): e2020GL091247, 2021. doi: https://doi.org/10.1029/2020GL091247. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL091247. e2020GL091247 2020GL091247.

Janina Klatt, Hans Peter Schmid, Matthias Mauder, and Rainer Steinbrecher. Fluxnet2015 DE-SfN Schechenfilz Nord. URL https://doi.org/gr2s8t.

Alexander Knohl, Frank Tiedemann, Olaf Kolle, Ernst-Detlef Schulze, Peter Anthoni, Werner Kutsch, Mathias Herbst, and Lukas Siebicke. Fluxnet2015 DE-Lnf Leinefelde. URL https://doi.org/gr2s65.

Sara Knox. Ameriflux FLUXNET-1F CA-DB2 Delta Burns Bog 2. URL https://doi.org/gr2sz8.

Hideki Kobayashi and Rikie Suzuki. Fluxnet2015 US-Prr Poker Flat Research Range Black Spruce Forest. URL https://doi.org/gr2s6b.

Ch. Körner. *Leaf Diffusive Conductances in the Major Vegetation Types of the Globe*, pp. 463–490. Springer Berlin Heidelberg, Berlin, Heidelberg, 1995. ISBN 978-3-642-79354-7. doi: 10.1007/978-3-642-79354-7_22. URL https://doi.org/10.1007/978-3-642-79354-7_22.

Yoshiko Kosugi and Satoru Takanashi. Fluxnet2015 MY-PSO Pasoh Forest Reserve (PSO). URL https://doi.org/gr2s9h.

Ayumi Kotani. Fluxnet2015 JP-SMF Seto Mixed Forest Site. URL https://doi.org/gr2s9g.

Shirley Kurc. Fluxnet2015 US-SRC Santa Rita Creosote. URL https://doi.org/hchx8z.

Gitta Lasslop, Markus Reichstein, Dario Papale, Andrew D. Richardson, Almut Arneth, Alan Barr, Paul Stoy, and Georg Wohlfahrt. Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. *Global Change Biology*, 16(1):187–208, 2010. doi: https://doi.org/10.1111/j.1365-2486.2009.02041.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2009.02041.x.

Bev Law. Fluxnet2015 US-Me6 Metolius Young Pine Burn. URL https://doi.org/gr2s52.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

Joel Leonard, Frédéric Bornet, and Eric Grehan. Etc l2 fluxnet (half-hourly) from estreesmons a28, 2016-12-31–2021-12-31, 2022. URL https://hdl.handle.net/11676/2-2NwQ92D9mPkNJkkrXpUs-t.

Qingliang Li, Cheng Zhang, Wei Shangguan, Zhongwang Wei, Hua Yuan, Jinlong Zhu, Xiaoning Li, Lu Li, Gan Li, Pingping Liu, and Yongjiu Dai. Landbench 1.0: A benchmark dataset and evaluation metrics for data-driven land surface variables prediction. *Expert Systems with Applications*, 243: 122917, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.122917. URL https://www.sciencedirect.com/science/article/pii/S095741742303419X.

Yingnian Li. Fluxnet2015 CN-Ha2 Haibei Shrubland. URL https://doi.org/hchx7v.

Michael J. Liddell. Fluxnet2015 AU-Rob Robson Creek, Queensland, Australia. URL https://doi.org/gr2s8f.

Matthias Lindauer, Rainer Steinbrecher, Benjamin Wolpert, Matthias Mauder, and Hans Peter Schmid. Fluxnet2015 DE-Lkb Lackenberg. URL https://doi.org/gr2s8r.

John Lloyd and JA Taylor. On the temperature dependence of soil respiration. *Functional ecology*, pp. 315–323, 1994.

Siyan Ma, Liukang Xu, Joseph Verfaillie, and Dennis Baldocchi. Ameriflux FLUXNET-1F US-Var Vaira Ranch- Ione. URL https://doi.org/hchx84.

Craig Macfarlane, Patricia Lambert, John Byrne, Chris Johnstone, and Natalie Smart. Fluxnet2015 AU-Gin Gingin, a. URL https://doi.org/gr2s8c.

Craig Macfarlane, Suzanne Prober, and Georg Wiehl. Fluxnet2015 AU-GWW Great Western Woodlands, Western Australia, Australia, b. URL https://doi.org/hchx7k.

Miguel D. Mahecha, Markus Reichstein, Nuno Carvalhais, Gitta Lasslop, Holger Lange, Sonia I. Seneviratne, Rodrigo Vargas, Christof Ammann, M. Altaf Arain, Alessandro Cescatti, Ivan A. Janssens, Mirco Migliavacca, Leonardo Montagnani, and Andrew D. Richardson. Global convergence in the temperature sensitivity of respiration at ecosystem level. *Science*, 329(5993):838–840, 2010. doi: 10.1126/science.1189587. URL https://www.science.org/doi/abs/10.1126/science.1189587.

Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

Hank A. Margolis. Fluxnet2015 CA-Qfo Quebec - Eastern Boreal, Mature Black Spruce. URL https://doi.org/gr2s37.

Giorgio Matteucci. Fluxnet2015 IT-Col Collelongo. URL https://doi.org/gr2s7g.

Harry McCaughey. Fluxnet2015 CA-Gro Ontario - Groundhog River, Boreal Mixedwood Forest. URL https://doi.org/gr2s3p.

Hanna Meyer and Edzer Pebesma. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1):2208, Apr 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29838-9. URL https://doi.org/10.1038/s41467-022-29838-9.

Wayne Meyer, Peter Cale, Georgia Koerber, Cacilia Ewenz, and Qiaoqi Sun. Fluxnet2015 AU-Cpr Calperum. URL https://doi.org/gr2s77.

Tilden Meyers. Fluxnet2015 US-Goo Goodwin Creek. URL https://doi.org/gr2s43.

Antje M. Moffat, Dario Papale, Markus Reichstein, David Y. Hollinger, Andrew D. Richardson, Alan G. Barr, Clemens Beckstein, Bobby H. Braswell, Galina Churkina, Ankur R. Desai, Eva Falge, Jeffrey H. Gove, Martin Heimann, Dafeng Hui, Andrew J. Jarvis, Jens Kattge, Asko Noormets, and Vanessa J. Stauch. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, 147(3):209–232, 2007. ISSN 0168-1923. doi: https://doi.org/10.1016/j.agrformet.2007.08.011. URL https://www.sciencedirect.com/science/article/pii/S016819230700216X.

Talie Musavi, Mirco Migliavacca, Markus Reichstein, Jens Kattge, Christian Wirth, T. Andrew Black, Ivan Janssens, Alexander Knohl, Denis Loustau, Olivier Roupsard, Andrej Varlagin, Serge Rambal, Alessandro Cescatti, Damiano Gianelle, Hiroaki Kondo, Rijan Tamrakar, and Miguel D. Mahecha. Stand age and species richness dampen interannual variation of ecosystem-level photosynthetic capacity. *Nature Ecology & Evolution*, 1(2):0048, Jan 2017. ISSN 2397-334X. doi: 10.1038/s41559-016-0048. URL https://doi.org/10.1038/s41559-016-0048.

J. A. Nelson, S. Walther, F. Gans, B. Kraft, U. Weber, K. Novick, N. Buchmann, M. Migliavacca, G. Wohlfahrt, L. Šigut, A. Ibrom, D. Papale, M. Göckede, G. Duveiller, A. Knohl, L. Hörtnagl, R. L. Scott, W. Zhang, Z. M. Hamdi, M. Reichstein, S. Aranda-Barranco, J. Ardö, M. Op de Beeck, D. Billdesbach, D. Bowling, R. Bracho, C. Brümmer, G. Camps-Valls, S. Chen, J. R. Cleverly, A. Desai, G. Dong, T. S. El-Madany, E. S. Euskirchen, I. Feigenwinter, M. Galvagno, G. Gerosa, B. Gielen, I. Goded, S. Goslee, C. M. Gough, B. Heinesch, K. Ichii, M. A. Jackowicz-Korczynski, A. Klosterhalfen, S. Knox, H. Kobayashi, K.-M. Kohonen, M. Korkiakoski, I. Mammarella, G. Mana, R. Marzuoli, R. Matamala, S. Metzger, L. Montagnani, G. Nicolini, T. O'Halloran, J.-M. Ourcival, M. Peichl, E. Pendall, B. Ruiz Reverter, M. Roland, S. Sabbatini, T. Sachs, M. Schmidt, C. R. Schwalm, A. Shekhar, R. Silberstein, M. L. Silveira, D. Spano, T. Tagesson, G. Tramontana, C. Trotta, F. Turco, T. Vesala, C. Vincke, D. Vitale, E. R. Vivoni, Y. Wang, W. Woodgate, E. A. Yepez, J. Zhang, D. Zona, and M. Jung. X-base: the first terrestrial carbon and water flux products from an extended data-driven scaling framework, fluxcom-x. *EGUsphere*, 2024:1–51, 2024. doi: 10.5194/egusphere-2024-165. URL https://egusphere.copernicus.org/preprints/2024/egusphere-2024-165/.

NEON Network). Ameriflux FLUXNET-1F US-xSR NEON Santa Rita Experimental Range (SRER), a. URL https://doi.org/hchx96.

NEON Network). Ameriflux FLUXNET-1F US-xNQ NEON Onaqui-Ault (ONAQ), b. URL https://doi.org/hchx9r.

NEON Network). Ameriflux FLUXNET-1F US-xSB NEON Ordway-Swisher Biological Station (OSBS), c. URL https://doi.org/hchx94.

NEON Network). Ameriflux FLUXNET-1F US-xHA NEON Harvard Forest (HARV), d. URL https://doi.org/hchx9g.

NEON Network). Ameriflux FLUXNET-1F US-xCP NEON Central Plains Experimental Range (CPER), e. URL https://doi.org/hchx89.

NEON Network). Ameriflux FLUXNET-1F US-xJE NEON Jones Ecological Research Center (JERC), f. URL https://doi.org/hchx9j.

NEON Network). Ameriflux FLUXNET-1F US-xCL NEON LBJ National Grassland (CLBJ), g. URL https://doi.org/hchx88.

NEON Network). Ameriflux FLUXNET-1F US-xNG NEON Northern Great Plains Research Laboratory (NOGP), h. URL https://doi.org/hchx9q.

NEON Network). Ameriflux FLUXNET-1F US-xKZ NEON Konza Prairie Biological Station (KONZ), i. URL https://doi.org/hchx9m.

NEON Network). Ameriflux FLUXNET-1F US-xTR NEON Treehaven (TREE), j. URL https://doi.org/hchx98.

NEON Network). Ameriflux FLUXNET-1F US-xGR NEON Great Smoky Mountains National Park, Twin Creeks (GRSM), k. URL https://doi.org/hchx9f.

NEON Network). Ameriflux FLUXNET-1F US-xYE NEON Yellowstone Northern Range (Frog Rock) (YELL), l. URL https://doi.org/hchzbc.

NEON Network). Ameriflux FLUXNET-1F US-xDS NEON Disney Wilderness Preserve (DSNY), m. URL https://doi.org/hchx9d.

NEON Network). Ameriflux FLUXNET-1F US-xST NEON Steigerwaldt Land Services (STEI), n. URL https://doi.org/hchx97.

NEON Network). Ameriflux FLUXNET-1F US-xAE NEON Klemme Range Research Station (OAES), o. URL https://doi.org/hchx87.

NEON Network). Ameriflux FLUXNET-1F US-xBR NEON Bartlett Experimental Forest (BART), p. URL https://doi.org/gr2s3k.

NEON Network). Ameriflux FLUXNET-1F US-xML NEON Mountain Lake Biological Station (MLBS), q. URL https://doi.org/hchx9p.

NEON Network). Ameriflux FLUXNET-1F US-xTA NEON Talladega National Forest (TALL), r. URL https://doi.org/hchx3s.

NEON Network). Ameriflux FLUXNET-1F US-xRM NEON Rocky Mountain National Park, CASTNET (RMNP), s. URL https://doi.org/hchx93.

NEON Network). Ameriflux FLUXNET-1F US-xKA NEON Konza Prairie Biological Station - Relocatable (KONA), t. URL https://doi.org/hchx9k.

NEON Network). Ameriflux FLUXNET-1F US-xUK NEON The University of Kansas Field Station (UKFS), u. URL https://doi.org/hchx99.

NEON Network). Ameriflux FLUXNET-1F US-xHE NEON Healy (HEAL), v. URL https://doi.org/hchx9h.

NEON Network). Ameriflux FLUXNET-1F US-xUN NEON University of Notre Dame Environmental Research Center (UNDE), w. URL https://doi.org/hchzbb.

NEON Network). Ameriflux FLUXNET-1F US-xSE NEON Smithsonian Environmental Research Center (SERC), x. URL https://doi.org/hchx95.

NEON Network). Ameriflux FLUXNET-1F US-xDL NEON Dead Lake (DELA), y. URL https://doi.org/hchx9c.

Ngoc B. Nguyen, Mirco Migliavacca, Maoya Bassiouni, Dennis D. Baldocchi, Laureano A. Gherardi, Julia K. Green, Dario Papale, Markus Reichstein, Kai-Hendrik Cohrs, Alessandro Cescatti, Tuan Dung Nguyen, Hoang H. Nguyen, Quang Minh Nguyen, and Trevor F. Keenan. Widespread underestimation of rain-induced soil carbon emissions from global drylands. *Nature Geoscience*, 18(9):869–876, Sep 2025. ISSN 1752-0908. doi: 10.1038/s41561-025-01754-9. URL https://doi.org/10.1038/s41561-025-01754-9.

Asko Noormets, Ge Sun, Michael Gavazzi, Steve McNulty, Jean-Christophe Domec, and John King. Ameriflux FLUXNET-1F US-NC1 NC_Clearcut. URL https://doi.org/hchx8w.

Ruben Ohana, Michael McCabe, Lucas Thibaut Meyer, Rudy Morel, Fruzsina Julia Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond Buschman Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich Kerswell, Suryanarayana Maddu, Jonah M. Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam Holden Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=00Sx577BT3.

Andrew Ouimette and Roel Ruzol. Ameriflux FLUXNET-1F US-Ho2 Howland Forest (west tower). URL https://doi.org/gr2s2x.

Jean-Marc Ourcival. Fluxnet2015 FR-Pue Puechabon. URL https://doi.org/gr2s7f.

Dario Papale, Giampiero Tirone, Riccardo Valentini, Nicola Arriga, Luca Belelli, Claudia Consalvo, Sabina Dore, Giovanni Manca, Francesco Mazzenga, Simone Sabbatini, and Paolo Stefani. Fluxnet2015 IT-Ro2 Roccarespampani 2. URL https://doi.org/gr2s7p.

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, July 2020.

Elise Pendall and Anne Griebel. Fluxnet2015 AU-Cum Cumberland Plains. URL `https://doi.org/gr2s78`.

Andrea Pitacco, Luca Tezza, Franco Meggio, Alessandro Peressotti, and Nadia Vendrame. Etc l2 fluxnet (half-hourly) from lison, 2015-12-31–2021-12-31, 2022. URL `https://hdl.handle.net/11676/5b1r2cK2QQxHcZBqbNiF3h__`.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. Dataset shift in machine learning. 2009.

M. Reichstein, G. Camps-Valls, B. Stevens, J. Denzler, N. Carvalhais, M. Jung, and Prabhat. Deep learning and process understanding for data-driven E arth system science. *Nature*, 566:195–204, Feb 2019.

Markus Reichstein, Eva Falge, Dennis Baldocchi, Dario Papale, Marc Aubinet, Paul Berbigier, Christian Bernhofer, Nina Buchmann, Tagir Gilmanov, Andre Granier, et al. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global change biology*, 11(9):1424–1439, 2005a.

Markus Reichstein, Eva Falge, Dennis Baldocchi, Dario Papale, Marc Aubinet, Paul Berbigier, Christian Bernhofer, Nina Buchmann, Tagir Gilmanov, André Granier, Thomas Grünwald, Katka Havránková, Hannu Ilvesniemi, Dalibor Janous, Alexander Knohl, Tuomas Laurila, Annalea Lohila, Denis Loustau, Giorgio Matteucci, Tilden Meyers, Franco Miglietta, Jean-Marc Ourcival, Jukka Pumpanen, Serge Rambal, Eyal Rotenberg, Maria Sanz, John Tenhunen, Günther Seufert, Francesco Vaccari, Timo Vesala, Dan Yakir, and Riccardo Valentini. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology*, 11(9):1424–1439, 2005b. doi: https://doi.org/10.1111/j.1365-2486.2005.001002.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2005.001002.x.

Camilo Rey-Sanchez, Carlos Wang, Daphne Szutu, Kyle Hemes, Joseph Verfaillie, and Dennis Baldocchi. Ameriflux FLUXNET-1F US-Bi2 Bouldin Island corn, a. URL https://doi.org/gr2szr.

Camilo Rey-Sanchez, Carlos Wang, Daphne Szutu, Robert Shortt, Samuel Chamberlain, Joseph Verfaillie, and Dennis Baldocchi. Ameriflux FLUXNET-1F US-Bi1 Bouldin Island Alfalfa, b. URL https://doi.org/gr2szq.

Cynthia Rudin and Joanna Radin. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2), nov 22 2019. https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

Simone Sabbatini, Nicola Arriga, Beniamino Gioli, and Dario Papale. Fluxnet2015 IT-CA2 Castel d'Asso2, a. URL https://doi.org/gr2s88.

Simone Sabbatini, Nicola Arriga, Giorgio Matteucci, and Dario Papale. Fluxnet2015 IT-CA3 Castel d'Asso 3, b. URL https://doi.org/gr2s89.

Simone Sabbatini, Nicola Arriga, and Dario Papale. Fluxnet2015 IT-CA1 Castel d'Asso1, c. URL https://doi.org/gr2s87.

Daniel Schermaier, Jeremy Forsythe, Michael Kline, and Thomas O'Halloran. Ameriflux FLUXNET-1F US-HB1 North Inlet Crab Haul Creek. URL https://doi.org/gr2sx4.

Marius Schmidt, Nils Becker, Daniel Dolfus, Alexander Graf, Martina Kettler, and Judith Mattes. Etc l2 fluxnet (half-hourly) from rollesbroich, 2010-12-31–2021-12-31, 2022. URL https://hdl.handle.net/11676/POOKwHoQty8EN0xRHoBbfx28.

Ivan Schroder, Steve Zegelin, Tehani Palu, and Andrew Feitz. Fluxnet2015 AU-Emr Emerald. URL https://doi.org/gr2s8b.

Russ Scott. Ameriflux FLUXNET-1F US-Whs Walnut Gulch Lucky Hills Shrub, a. URL https://doi.org/hchx85.

Russell Scott. Fluxnet2015 US-SRG Santa Rita Grassland, b. URL https://doi.org/gr2s6c.

Russell Scott. Fluxnet2015 US-SRM Santa Rita Mesquite, c. URL https://doi.org/gr2s5q.

Russell Scott. Ameriflux FLUXNET-1F US-Wkg Walnut Gulch Kendall Grasslands, d. URL https://doi.org/hchx86.

Peili Shi, Xianzhou Zhang, and Yongtao He. Fluxnet2015 CN-Dan Dangxiong. URL https://doi.org/gr2s6v.

Ladislav Sigut, Katerina Havrankova, Georg Jocher, Marian Pavelka, and Dalibor Janouš. Fluxnet2015 CZ-BK2 Bily Kriz grassland. URL https://doi.org/gr2s63.

Maria Silveira and Rosvel Bracho. Ameriflux FLUXNET-1F US-ONA Florida pine flatwoods. URL https://doi.org/gr2sx8.

Masashi Sugiyama and Motoaki Kawanabe. *Learning Under Covariate Shift*, pp. 19–19. 2012.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBENCH: AN EXTENSIVE BENCHMARK FOR SCI- ENTIFIC MACHINE LEARNING. In *ICLR 2023 Workshop on Physics for Machine Learning*, 2023. URL https://openreview.net/forum?id=b8SwOxZQ2kj.

Drought 2018 Team and ICOS Ecosystem Thematic Centre. Drought-2018 ecosystem eddy covariance flux product for 52 stations in FLUXNET-Archive format, a. URL https://doi.org/gr2s9r.

Warm Winter 2020 Team and ICOS Ecosystem Thematic Centre. Warm Winter 2020 ecosystem eddy covariance flux product for 73 stations in FLUXNET-Archive format—release 2022-1, b. URL https://doi.org/gr2s9n.

T. Thum, S. Caldararu, J. Engel, M. Kern, M. Pallandt, R. Schnur, L. Yu, and S. Zaehle. A new model of the coupled carbon, nitrogen, and phosphorus cycles in the terrestrial biosphere (quincy v1.0; revision 1996). *Geoscientific Model Development*, 12(11):4781–4802, 2019. doi: 10.5194/gmd-12-4781-2019. URL https://gmd.copernicus.org/articles/12/4781/2019/.

Gianluca Tramontana, Mirco Migliavacca, Martin Jung, Markus Reichstein, Trevor F. Keenan, Gustau Camps-Valls, Jerome Ogee, Jochem Verrelst, and Dario Papale. Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks. *Global Change Biology*, 26(9): 5235–5253, 2020. doi: https://doi.org/10.1111/gcb.15203. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15203.

Violeta Teodora Trifunov, Maha Shadaydeh, Jakob Runge, Markus Reichstein, and Joachim Denzler. A data-driven approach to partitioning net ecosystem exchange using a deep state space model. *IEEE Access*, 9:107873–107883, 2021. doi: 10.1109/ACCESS.2021.3101129.

Eeva-Stiina Tuittila, Pasi Kolari, Aino Korrensalo, Heikki Laakso, Janne Levula, Ivan Mammarella, Teemu Matilainen, Elisa Männistö, Risto Taipale, and Timo Vesala. Etc l2 fluxnet (half-hourly) from siikaneva, 2017-12-31–2022-12-31, 2023. URL https://hdl.handle.net/11676/QfuBdfWZTepx1GZ5ot4VOI5h.

Riccardo Valentini, Giacomo Nicolini, Paolo Stefani, Agnès De Grandcourt, and Silvio Stivanello. Fluxnet2015 GH-Ank Ankasa, a. URL https://doi.org/gr2s86.

Riccardo Valentini, Giampiero Tirone, Domenico Vitale, Dario Papale, Nicola Arriga, Luca Belelli, Sabina Dore, Giovanni Manca, Francesco Mazzenga, Emiliano Pegoraro, Simone Sabbatini, and Paolo Stefani. Fluxnet2015 IT-Ro1 Roccarespampani 1, b. URL https://doi.org/gr2s7n.

Jacobus Henricus Van't Hoff, Robert Alfred Lehfeldt, et al. Lectures on theoretical and physical chemistry. 1899.

Henriikka Vekuri, Juha-Pekka Tuovinen, Liisa Kulmala, Dario Papale, Pasi Kolari, Mika Aurela, Tuomas Laurila, Jari Liski, and Annalea Lohila. A widely-used eddy covariance gap-filling method creates systematic bias in carbon balance estimates. *Scientific Reports*, 13(1):1720, Jan 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-28827-2. URL https://doi.org/10.1038/s41598-023-28827-2.

George Vourlitis, Higo Dalmagro, José de S. Nogueira, Mark Johnson, and Paulo Arruda. Ameriflux FLUXNET-1F BR-Npw Northern Pantanal Wetland. URL https://doi.org/gr2sz7.

Huimin Wang and Xiaoli Fu. Fluxnet2015 CN-Qia Qianyanzhou. URL https://doi.org/gr2s6z.

Zhihao Wang, Lei Ma, George Hurtt, Xiaowei Jia, Yanhua Li, Ruohan Li, Zhili Li, Shuo Xu, and Yiqun Xie. Carbonglobe: A global-scale, multi-decade dataset and benchmark for carbon forecasting in forest ecosystems. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0261. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0b82662b6c32e887bb252a74d8cb2d5e-Paper-Datasets_and_Benchmarks_Track.pdf.

D. Watson-Parris, Y. Rao, D. Olivié, Ø. Seland, P. Nowack, G. Camps-Valls, P. Stier, S. Bouabid, M. Dewey, E. Fons, J. Gonzalez, P. Harder, K. Jeggle, J. Lenhardt, P. Manshausen, M. Novitasari, L. Ricard, and C. Roesch. Climatebench v1.0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10):e2021MS002954, 2022. doi: https://doi.org/10.1029/2021MS002954. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002954. e2021MS002954 2021MS002954.

Georg Wohlfahrt, Albin Hammerle, and Lukas Hörtnagl. Fluxnet2015 AT-Neu Neustift. URL https://doi.org/gr2s6k.

Jeffrey Wood and Lianhong Gu. Ameriflux FLUXNET-1F US-MOz Missouri Ozark Site. URL https://doi.org/gr2szn.

T. Wutzler, A. Lucas-Moffat, M. Migliavacca, J. Knauer, K. Sickel, L. Šigut, O. Menzer, and M. Reichstein. Basic and extensible post-processing of eddy covariance flux data with reddyproc. *Biogeosciences*, 15(16):5015–5030, 2018. doi: 10.5194/bg-15-5015-2018. URL https://bg.copernicus.org/articles/15/5015/2018/.

Enrico Yepez and Jaime Garatuza. Ameriflux FLUXNET-1F MX-Tes Tesopaco, secondary tropical dry forest. URL https://doi.org/gr2sxw.

Sungduk Yu, Zeyuan Hu, Akshay Subramaniam, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus C. Will, Gunnar Behrens, Julius J. M. Busecke, Nora Loose, Charles I Stern, Tom Beucler, Bryce Harrop, Helge Heuer, Benjamin R Hillman, Andrea Jenney, Nana Liu, Alistair White, Tian Zheng, Zhiming Kuang, Fiaz Ahmed, Elizabeth Barnes, Noah D. Brenowitz, Christopher Bretherton, Veronika Eyring, Savannah Ferretti, Nicholas Lutsko, Pierre Gentine, Stephan Mandt, J. David Neelin, Rose Yu, Laure Zanna, Nathan M. Urban, Janni Yuval, Ryan Abernathey, Pierre Baldi, Wayne Chuang, Yu Huang, Fernando Iglesias-Suarez, Sanket Jantre, Po-Lun Ma, Sara Shamekh, Guang Zhang, and Michael Pritchard. Climsim-online: A large multi-scale dataset and framework for hybrid physics-ml climate emulation. *Journal of Machine Learning Research*, 26(142):1–85, 2025. URL http://jmlr.org/papers/v26/24-1014.html.

Weiwei Zhan, Xi Yang, Youngryel Ryu, Benjamin Dechant, Yu Huang, Yves Goulas, Minseok Kang, and Pierre Gentine. Two for one: Partitioning CO2 fluxes and understanding the relationship between solar-induced chlorophyll fluorescence and gross primary productivity using machine learning. *Agricultural and Forest Meteorology*, 321:108980, 2022. ISSN 0168-1923. doi: https://doi.org/10.1016/j.agrformet.2022.108980. URL https://www.sciencedirect.com/science/article/pii/S0168192322001708.

Junhui Zhang and Shijie Han. Fluxnet2015 CN-Cha Changbaishan. URL https://doi.org/gr2s6t.

Guoyi Zhou and Junhua Yan. Fluxnet2015 CN-Din Dinghushan. URL https://doi.org/gr2s6w.

A Benchmark Design

This section details how the benchmark is constructed: site selection, the mechanistic simulations that provide ground truth, the rain-pulse event model, the reintroduction of measurement noise and data gaps, the available drivers, and the full evaluation protocol.

A.1 SITES

The initial candidate set of eddy-covariance sites was taken from (Nelson et al., 2024) and comprises the 224 sites for which half-hourly meteorological forcing (typically 1990–2021) is available. Sites were retained only when land-surface model output existed for *both* QUINCY and CABLE-POP: eight sites had been run with neither model, and a further 18 lacked a QUINCY simulation, leaving 198 sites. Evaluation years were then restricted to the overlap of these simulations with FLUXNET half-hourly observations; because the model records end in 2021, later FLUXNET years were discarded. A calendar year was included only if QUINCY, CABLE-POP and the measured record each contained more than 175 half-hourly timesteps with directly measured NEE ($\mathrm{NEE}_{\mathrm{QC}} = 0$), so that partitioning is compared on a common window with sufficient observed NEE. Finally, site-years in which either model produced identically zero (zero-variance) GPP—indicating that vegetation was never established—were excluded, because GPP skill metrics are then undefined and the simulation is not a meaningful partitioning target. This left 188 sites and 1403 site-years (1996–2021) for all analyses reported in this paper.

In table 2, we provide a summary, and in table 3, we provide a full list of all sites included in the dataset.

Table 2: Summary of benchmark sites by land cover type and event category.

Category	**Type**	**Count**
Land cover	CRO — Cropland	23
	CSH — Closed Shrubland	1
	DBF — Deciduous Broadleaf Forest	27
	EBF — Evergreen Broadleaf Forest	12
	ENF — Evergreen Needleleaf Forest	38
	GRA — Grassland	41
	MF — Mixed Forest	11
	OSH — Open Shrubland	7
	SAV — Savanna	8
	WET — Permanent Wetland	15
	WSA — Woody Savanna	5
Total sites		188
Total site-years		1403
Event sites	Rain-pulse (RP)	18

Table 3: Summary of the sites included in the dataset.

Site ID	**IGBP**	**Lat**	**Lon**	**Years**	**Event**	**Reference**
AT-Neu	GRA	47.1167	11.3175	11	–	Wohlfahrt et al.
AU-ASM	SAV	-22.2830	133.2490	5	RP	Cleverly & Eamus (b)
AU-Cpr	SAV	-34.0021	140.5891	5	RP	Meyer et al.
AU-Cum	EBF	-33.6152	150.7236	3	–	Pendall & Griebel
AU-DaP	GRA	-14.0633	131.3181	7	RP	Beringer & Hutley (c)
AU-DaS	SAV	-14.1593	131.3881	7	RP	Beringer & Hutley (f)
AU-Dry	SAV	-15.2588	132.3706	7	RP	Beringer & Hutley (e)
AU-Emr	GRA	-23.8587	148.4746	3	–	Schroder et al.
AU-GWW	SAV	-30.1913	120.6541	2	–	Macfarlane et al. (b)
AU-Gin	WSA	-31.3764	115.7138	4	RP	Macfarlane et al. (a)
AU-How	WSA	-12.4943	131.1523	14	–	Beringer & Hutley (b)
AU-Lox	DBF	-34.4704	140.6551	2	–	Ewenz et al.
AU-RDF	WSA	-14.5636	132.4776	3	–	Beringer & Hutley (d)
AU-Rig	GRA	-36.6499	145.5759	4	RP	Beringer et al. (a)
AU-Rob	EBF	-17.1175	145.6301	1	–	Liddell

continued on next page

(continued from previous page)

Site ID	IGBP	Lat	Lon	Years	Event	Reference
AU-Stp	GRA	-17.1507	133.3502	7	RP	Beringer & Hutley (a)
AU-TTE	GRA	-22.2870	133.6400	3	–	Cleverly & Eamus (a)
AU-Wac	EBF	-37.4259	145.1878	4	–	Beringer et al. (b)
AU-Wom	EBF	-37.4222	144.0944	5	–	Arndt et al.
AU-Ync	GRA	-34.9893	146.2907	3	–	Beringer & Walker
BE-Bra	MF	51.3076	4.5198	15	–	Team & Centre (b)
BE-Dor	GRA	50.3119	4.9680	10	–	Team & Centre (b)
BE-Lon	CRO	50.5516	4.7462	11	–	Team & Centre (b)
BE-Maa	CSH	50.9799	5.6319	5	–	Team & Centre (b)
BE-Vie	MF	50.3049	5.9981	19	–	Team & Centre (b)
BR-Npw	WSA	-16.4980	-56.4120	5	–	Vourlitis et al.
BR-Sa3	EBF	-3.0180	-54.9714	5	–	Goulden
CA-DB2	WET	49.1190	-122.9951	2	–	Knox
CA-DBB	WET	49.1293	-122.9849	7	–	Christen & Knox
CA-Gro	MF	48.2167	-82.1556	12	–	McCaughey
CA-LP1	ENF	55.1119	-122.8414	12	–	Black (c)
CA-Oas	DBF	53.6289	-106.1978	15	–	Black (b)
CA-Obs	ENF	53.9872	-105.1178	12	–	Black (a)
CA-Qfo	ENF	49.6925	-74.3421	8	–	Margolis
CA-SF1	ENF	54.4850	-105.8176	4	–	Amiro (b)
CA-SF2	ENF	54.2539	-105.8775	5	–	Amiro (a)
CA-SF3	OSH	54.0916	-106.0053	6	–	Amiro (c)
CA-TP4	ENF	42.7102	-80.3574	13	–	Arain (b)
CA-TPD	DBF	42.6353	-80.5577	3	–	Arain (a)
CH-Aws	GRA	46.5831	9.7906	14	–	Team & Centre (b)
CH-Cha	GRA	47.2102	8.4104	10	–	Team & Centre (b)
CH-Dav	ENF	46.8153	9.8559	18	–	Team & Centre (b)
CH-Fru	GRA	47.1158	8.5378	10	–	Team & Centre (b)
CH-Lae	MF	47.4783	8.3644	11	–	Team & Centre (b)
CH-Oe1	GRA	47.2858	7.7319	7	–	Ammann
CH-Oe2	CRO	47.2864	7.7337	17	–	Team & Centre (b)
CN-Cha	MF	42.4025	128.0958	3	–	Zhang & Han
CN-Cng	GRA	44.5934	123.5092	4	–	Dong
CN-Dan	GRA	30.4978	91.0664	2	–	Shi et al.
CN-Din	EBF	23.1733	112.5361	3	–	Zhou & Yan
CN-Ha2	WET	37.6086	101.3269	3	–	Li
CN-Qia	ENF	26.7414	115.0581	3	–	Wang & Fu
CZ-BK1	ENF	49.5021	18.5369	17	–	Team & Centre (b)
CZ-BK2	GRA	49.4944	18.5429	7	–	Sigut et al.
CZ-KrP	CRO	49.5733	15.0788	7	–	Team & Centre (b)
CZ-Lnz	MF	48.6815	16.9463	6	–	Team & Centre (b)
CZ-RAJ	ENF	49.4437	16.6965	9	–	Team & Centre (b)
CZ-Stn	DBF	49.0360	17.9699	11	–	Team & Centre (b)
CZ-wet	WET	49.0247	14.7704	15	–	Team & Centre (b)
DE-Geb	CRO	51.0997	10.9146	14	–	Team & Centre (b)
DE-Gri	GRA	50.9500	13.5126	11	–	Team & Centre (b)
DE-Hai	DBF	51.0792	10.4522	19	–	Team & Centre (b)
DE-HoH	DBF	52.0866	11.2224	6	–	Team & Centre (b)
DE-Hzd	DBF	50.9638	13.4898	11	–	Team & Centre (b)
DE-Kli	CRO	50.8931	13.5224	11	–	Team & Centre (b)
DE-Lkb	ENF	49.0996	13.3047	5	–	Lindauer et al.
DE-Lnf	DBF	51.3282	10.3678	8	–	Knohl et al.
DE-Obe	ENF	50.7867	13.7213	7	–	Team & Centre (b)
DE-RuR	GRA	50.6219	6.3041	4	–	Schmidt et al. (2022)
DE-RuS	CRO	50.8659	6.4471	4	–	Team & Centre (b)

continued on next page

(continued from previous page)

Site ID	IGBP	Lat	Lon	Years	Event	Reference
DE-RuW	ENF	50.5049	6.3310	9	–	Team & Centre (a)
DE-SfN	WET	47.8064	11.3275	3	–	Klatt et al.
DE-Tha	ENF	50.9626	13.5651	19	–	Team & Centre (b)
DK-Sor	DBF	55.4859	11.6446	19	–	Team & Centre (b)
DK-Vng	CRO	56.0375	9.1607	2	–	Friborg et al. (2023)
ES-Abr	SAV	38.7018	-6.7859	6	RP	Team & Centre (b)
ES-LM1	SAV	39.9427	-5.7787	7	RP	Team & Centre (b)
ES-LM2	SAV	39.9346	-5.7759	7	RP	Team & Centre (b)
FI-Hyy	ENF	61.8474	24.2948	19	–	Team & Centre (b)
FI-Let	ENF	60.6418	23.9595	12	–	Team & Centre (b)
FI-Lom	WET	67.9972	24.2092	3	–	Aurela et al.
FI-Sii	WET	61.8326	24.1928	5	–	Tuittila et al. (2023)
FR-Aur	CRO	43.5496	1.1061	16	–	Team & Centre (b)
FR-EM2	CRO	49.8721	3.0206	5	–	Leonard et al. (2022)
FR-FBn	MF	43.2408	5.6787	13	–	Team & Centre (b)
FR-Gri	CRO	48.8442	1.9519	11	–	Team & Centre (b)
FR-Hes	DBF	48.6741	7.0646	7	–	Team & Centre (b)
FR-LBr	ENF	44.7171	-0.7693	13	–	Berbigier & Loustau
FR-LGt	WET	47.3229	2.2841	4	–	Jacotot et al. (2022)
FR-Lam	CRO	43.4964	1.2379	16	–	Team & Centre (b)
FR-Mej	GRA	48.1184	-1.7964	3	–	Flechard et al. (2022)
FR-Pue	EBF	43.7413	3.5957	15	–	Ourcival
FR-Tou	GRA	43.5729	1.3747	3	–	Calvet et al. (2022)
GF-Guy	EBF	5.2788	-52.9249	11	–	Team & Centre (b)
GH-Ank	EBF	5.2685	-2.6942	3	–	Valentini et al. (a)
IT-BCi	CRO	40.5237	14.9574	11	–	Team & Centre (b)
IT-CA1	DBF	42.3804	12.0266	4	–	Sabbatini et al. (c)
IT-CA2	CRO	42.3772	12.0260	4	–	Sabbatini et al. (a)
IT-CA3	DBF	42.3800	12.0222	4	–	Sabbatini et al. (b)
IT-Col	DBF	41.8494	13.5881	19	–	Matteucci
IT-Cp2	EBF	41.7043	12.3573	3	–	Team & Centre (b)
IT-La2	ENF	45.9542	11.2853	3	–	Cescatti et al.
IT-Lav	ENF	45.9562	11.2813	12	–	Team & Centre (b)
IT-Lsn	OSH	45.7405	12.7503	5	–	Pitacco et al. (2022)
IT-MBo	GRA	46.0147	11.0458	11	–	Team & Centre (b)
IT-Ren	ENF	46.5869	11.4337	12	–	Team & Centre (b)
IT-Ro1	DBF	42.4081	11.9300	9	–	Valentini et al. (b)
IT-Ro2	DBF	42.3903	11.9209	2	–	Papale et al.
IT-Tor	GRA	45.8444	7.5781	7	–	Team & Centre (b)
JP-SMF	MF	35.2617	137.0788	5	–	Kotani
MX-Tes	DBF	27.8446	-109.2977	5	–	Yepez & Garatuza
MY-PSO	EBF	2.9730	102.3062	7	–	Kosugi & Takanashi
NL-Hor	GRA	52.2403	5.0713	8	–	Dolman et al.
NL-Loo	ENF	52.1666	5.7436	19	–	Team & Centre (a)
RU-Fy2	ENF	56.4476	32.9019	6	–	Team & Centre (b)
RU-Fyo	ENF	56.4615	32.9221	17	–	Team & Centre (b)
SE-Deg	WET	64.1820	19.5565	20	–	Team & Centre (b)
SE-Htm	ENF	56.0976	13.4190	6	–	Team & Centre (b)
SE-Nor	EBF	60.0865	17.4795	7	–	Team & Centre (b)
SE-Svb	ENF	64.2561	19.7745	6	–	Team & Centre (b)
US-A32	GRA	36.8193	-97.8198	3	–	Billesbach et al. (c)
US-AR1	GRA	36.4267	-99.4200	4	–	Billesbach et al. (b)
US-AR2	GRA	36.6358	-99.5975	4	–	Billesbach et al. (a)
US-ARM	CRO	36.6058	-97.4888	10	–	Biraud et al.
US-BZB	WET	64.6955	-148.3208	10	–	Euskirchen (a)

continued on next page

(continued from previous page)

Site ID	IGBP	Lat	Lon	Years	Event	Reference
US-BZF	WET	64.7013	-148.3121	10	–	Euskirchen (b)
US-Bi1	CRO	38.0992	-121.4993	6	–	Rey-Sanchez et al. (b)
US-Bi2	CRO	38.1091	-121.5351	5	–	Rey-Sanchez et al. (a)
US-CRT	CRO	41.6285	-83.3471	3	–	Chen & Chu (b)
US-CS1	CRO	44.1031	-89.5379	2	–	Desai (a)
US-CS3	CRO	44.1394	-89.5727	2	–	Desai (d)
US-CS4	CRO	44.1597	-89.5475	2	–	Desai (b)
US-GLE	ENF	41.3665	-106.2399	10	–	Frank et al.
US-Goo	GRA	34.2547	-89.8735	5	–	Meyers
US-HB1	WET	33.3455	-79.1957	3	–	Schermaier et al.
US-HWB	GRA	40.8608	-77.8488	4	–	Goslee
US-Ho2	ENF	45.2091	-68.7470	23	–	Ouimette & Ruzol
US-KFS	GRA	39.0561	-95.1907	12	–	Brunsell (a)
US-KLS	GRA	38.7745	-97.5684	8	–	Brunsell (b)
US-KS3	WET	28.7084	-80.7427	1	–	Bracho & Hinkle
US-Los	WET	46.0827	-89.9792	11	–	Desai (f)
US-MOz	DBF	38.7441	-92.2000	18	–	Wood & Gu
US-Me2	ENF	44.4523	-121.5574	13	–	Hanson et al.
US-Me3	ENF	44.3154	-121.6078	6	–	McCaughey
US-Me6	ENF	44.3233	-121.6078	5	–	Law
US-NC1	ENF	35.8118	-76.7119	8	–	Noormets et al.
US-NR1	ENF	40.0329	-105.5464	16	–	Blanken et al.
US-ONA	GRA	27.3836	-81.9509	6	–	Silveira & Bracho
US-Oho	DBF	41.5545	-83.8438	10	–	Chen et al.
US-Prr	ENF	65.1237	-147.4876	5	–	Kobayashi & Suzuki
US-Ro1	CRO	44.7143	-93.0898	13	–	Baker et al.
US-Ro4	GRA	44.6781	-93.0723	8	–	Baker & Griffis (a)
US-Ro5	CRO	44.6910	-93.0576	5	–	Baker & Griffis (c)
US-Ro6	CRO	44.6946	-93.0578	5	–	Baker & Griffis (b)
US-SRC	OSH	31.9083	-110.8395	7	RP	Kurc
US-SRG	GRA	31.7894	-110.8277	7	RP	Scott (b)
US-SRM	WSA	31.8214	-110.8661	11	RP	Scott (c)
US-Syv	MF	46.2420	-89.3477	10	–	Desai (e)
US-Var	GRA	38.4133	-120.9508	15	RP	Ma et al.
US-WCr	DBF	45.8059	-90.0799	13	–	Desai (c)
US-WPT	WET	41.4646	-82.9962	3	–	Chen & Chu (a)
US-Whs	OSH	31.7438	-110.0522	8	RP	Scott (a)
US-Wkg	GRA	31.7365	-109.9419	11	RP	Scott (d)
US-xAE	GRA	35.4106	-99.0588	3	–	Network) (o)
US-xBR	DBF	44.0639	-71.2873	3	–	Network) (p)
US-xCL	GRA	33.4012	-97.5700	3	–	Network) (g)
US-xCP	GRA	40.8155	-104.7456	3	–	Network) (e)
US-xDL	MF	32.5417	-87.8039	3	–	Network) (y)
US-xDS	GRA	28.1250	-81.4362	3	–	Network) (m)
US-xGR	DBF	35.6890	-83.5019	3	–	Network) (k)
US-xHA	DBF	42.5369	-72.1727	3	–	Network) (d)
US-xHE	OSH	63.8757	-149.2133	3	–	Network) (v)
US-xJE	ENF	31.1948	-84.4686	3	–	Network) (f)
US-xKA	GRA	39.1104	-96.6129	2	–	Network) (t)
US-xKZ	GRA	39.1008	-96.5631	3	–	Network) (i)
US-xML	DBF	37.3783	-80.5248	3	–	Network) (q)
US-xNG	GRA	46.7697	-100.9154	3	RP	Network) (h)
US-xNQ	OSH	40.1776	-112.4524	3	–	Network) (b)
US-xRM	ENF	40.2759	-105.5459	3	–	Network) (s)
US-xSB	ENF	29.6893	-81.9934	3	–	Network) (c)

continued on next page

(continued from previous page)

Site ID	IGBP	Lat	Lon	Years	Event	Reference
US-xSE	DBF	38.8901	-76.5600	3	–	Network) (x)
US-xSR	OSH	31.9107	-110.8355	3	–	Network) (a)
US-xST	DBF	45.5089	-89.5864	3	–	Network) (n)
US-xTA	ENF	32.9505	-87.3933	3	–	Network) (r)
US-xTR	DBF	45.4937	-89.5857	3	–	Network) (j)
US-xUK	DBF	39.0404	-95.1921	2	–	Network) (u)
US-xUN	MF	46.2339	-89.5373	3	–	Network) (w)
US-xYE	ENF	44.9535	-110.5391	3	–	Network) (l)

A.2 Mechanistic Simulations

The meteorological inputs are used to run two mechanistic land-surface models, QUINCY (Thum et al., 2019) and CABLE-POP (Haverd et al., 2018), each providing a time series of NEE and the corresponding ground-truth fluxes GPP and $\mathrm{R_{eco}}$. While both models simulate coupled carbon and nutrient cycles, they are structurally distinct in their core architectures, representations of vegetation dynamics, and photosynthesis. Generally, QUINCY focuses on a high-detail integration of the carbon, nitrogen, and phosphorus cycles, whereas CABLE-POP emphasizes the impacts of woody vegetation demography and land-use change. Regarding vegetation dynamics, QUINCY explicitly decouples carbon acquisition (source processes) from actual tissue formation (sink processes) to more realistically capture growth patterns under environmental stress. In contrast, CABLE-POP simulates cohorts of trees and representative patches of different ages, which allows for a highly realistic representation of biomass accumulation and stand structure over time.

A.3 Event Design: Rain-induced Respiration Pulses.

The pulse is applied to $\mathrm{R_{eco}}$ as an additive perturbation. For each daily time step i after the onset of a pulse event, the modified respiration is given by

$$\mathrm{R_{eco}^{pulse}}(t+i) = \mathrm{R_{eco}}(t+i) + \alpha \cdot 0.96 \cdot \exp(-0.16 \cdot i), \tag{2}$$

where α denotes the pulse intensity and i indexes daily time steps since the start of the event. The pulse is applied for up to 20 daily time steps to ensure a gradual fade-out without sudden drops. We then set the resulting daily pulse signal as midday values, which we interpolate to obtain a smooth, half-hourly decaying time series. Substituting the normal interpolated midday values with the perturbed midday values gives us the final half-hourly respiration time series.

NEE is finally recomputed as

$$\mathrm{NEE^{pulse}} = -\mathrm{GPP} + \mathrm{R_{eco}^{pulse}}. \tag{3}$$

We obtain information on event day and intensity from fluxtower annotations in Nguyen et al. (2025). Pulse onset is aligned to midday conditions by assigning events to a shifted daily time index, ensuring that the perturbation affects the diurnal cycle consistently with observed rain events. Fig fig. 8 shows an example of a rain-induced respiration pulse event with the fluxes before and after the artificial perturbation.

A.4 Reintroduction of Measurement Effects and Data Gaps

To mimic measurement uncertainty, noise is added to the simulated NEE. For each time step t, the observed flux $\mathrm{NEE_{obs},}_t$ is defined as

$$\mathrm{NEE_{obs},}_t = \mathrm{NEE}_t + \varepsilon_t, \tag{4}$$

where ε_t is sampled from a Laplace distribution, which captures the heavy-tailed error characteristics observed in EC measurements (Hollinger & Richardson, 2005), with zero mean and scale parameter b_t.

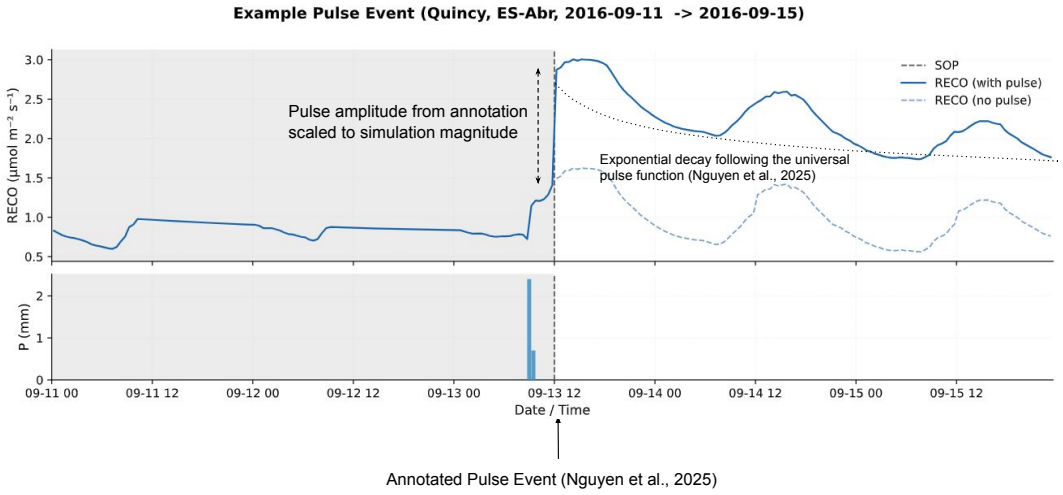


Figure 8: Example rain-induced respiration pulse event (Quincy, site ES-Abr, 2016-09-11 to 2016-09-15). Top: simulated $\mathrm{R_{eco}}$ with the imposed pulse (solid) versus the unperturbed baseline (dashed light blue). Following eq. (2), the pulse enters as a step increase at start of pulse (SOP) and fades over up to 20 daily steps. Pulse timing and amplitude follow site-specific annotations from Nguyen et al. [2025], scaled to the simulated flux magnitude. Bottom: half-hourly precipitation from the same driving meteorology, shown for verification: a precipitation event coincides with the annotated onset, though pulse magnitude is set from the external annotation rather than derived from this precipitation value directly.

The scale parameter b_t is derived from the site-specific random uncertainty estimate and rescaled to account for differences between simulated and observed flux magnitudes:

$$b_t = \frac{\sigma_t}{\sqrt{2}} \cdot s_t, \tag{5}$$

with

$$s_t = \frac{\max\left(|\mathrm{NEE}_t|, 1\right)}{\max\left(|\mathrm{NEE}_t^{\mathrm{meas}}|, 1\right)}. \tag{6}$$

Here, σ_t and $\mathrm{NEE}_t^{\mathrm{meas}}$ correspond to the variables `NEE_VUT_REF_RANDUNC` and `NEE_VUT_REF` from the FLUXNET dataset. Negative uncertainty values are clipped to zero.

In addition, we impose data gaps using measurement masks from FLUXNET, based on the quality flag `NEE_VUT_REF_QC`, to reflect realistic patterns of data availability. This step transforms the simulated fluxes into signals that exhibit the variability, noise structure, and missing-data characteristics for a specific flux-tower location.

A.5 Available Drivers

Table 4 lists the input variables available to the partitioning methods, following FLUXNET naming conventions. Individual methods draw only on the subsets specified in Table 8.

A.6 Data Coverage

The fraction of directly measured (non-gap-filled) NEE varies widely across site-years. Figure 9 shows this distribution, which motivates the coverage-stratified analysis reported in the main text.

A.7 Details of Evaluation Protocol

This section gives the precise definitions underlying the evaluation protocol. The temporal aggregation scales are defined in table 5, the stress-event subsets in table 6, and the error metrics in table 7.

Table 4: Input variables available to partitioning methods, following FLUXNET conventions.

Variable	Full name	Unit
Meteorological drivers		
TA	Air temperature	°C
SW_IN	Incoming shortwave radiation	$\mathrm{W\,m^{-2}}$
LW_IN	Incoming longwave radiation	$\mathrm{W\,m^{-2}}$
PA	Atmospheric pressure	kPa
RH	Relative humidity	%
VPD	Vapour pressure deficit	kPa
WS	Wind speed	$\mathrm{m\,s^{-1}}$
SW_IN_POT	Potential shortwave radiation	$\mathrm{W\,m^{-2}}$
SWC	Soil water content	$\mathrm{m^3\,m^{-3}}$
P	Precipitation	mm per 30 min
Energy balance		
LE	Latent heat flux	$\mathrm{W\,m^{-2}}$
H	Sensible heat flux	$\mathrm{W\,m^{-2}}$
G	Soil heat flux	$\mathrm{W\,m^{-2}}$
NETRAD	Net radiation	$\mathrm{W\,m^{-2}}$
Auxiliary		
NIGHT	Nighttime flag	Boolean
Target and quality		
NEE	Net ecosystem exchange	$\mu\mathrm{mol\,m^{-2}\,s^{-1}}$
USTAR	Friction velocity	$\mathrm{m\,s^{-1}}$
NEE_QC	NEE quality control flag	—

A.8 Using the Benchmark: Splits and Hold-out

Training, Validation, and Test Splits. It is standard practice to split data into train/validation/test folds. Here, however, it is not obvious what such a split would mean: a method usually consumes the entire observed NEE for a site and year to infer the decomposition, so there is no held-out portion in the supervised sense. Internal splitting is still relevant, e.g. for model selection or early stopping, but it interacts with dynamic events that may fall into the validation interval, letting the method fail to track them. Strategies such as k-fold with ensembling or checkerboard splitting could trade this off differently. Finally, whether to train only on measured points, on gap-filled ones, or to gap-fill differently is itself a choice we leave to the user.

Hold-out Scenarios. For the same reason, holding out a site or specific years is generally not meaningful. Cross-site or cross-year hold-out becomes meaningful for methods that transfer across locations or time, such as pretraining variants. We treat such splits as method-dependent and do not impose them.

B Baseline Methods

We include a variety of partitioning methods in the dataset, including parametric and machine learning methods. In table 8 we provide a summary of the partitioning methods included in the dataset.

B.1 Parametric Baselines

As parametric baselines, we include the nighttime-based partitioning method proposed by Reichstein et al. (2005b) and the daytime-based partitioning method proposed by Lasslop et al. (2010) as well as a modified daytime-based partitioning method proposed by Keenan et al. (2019).

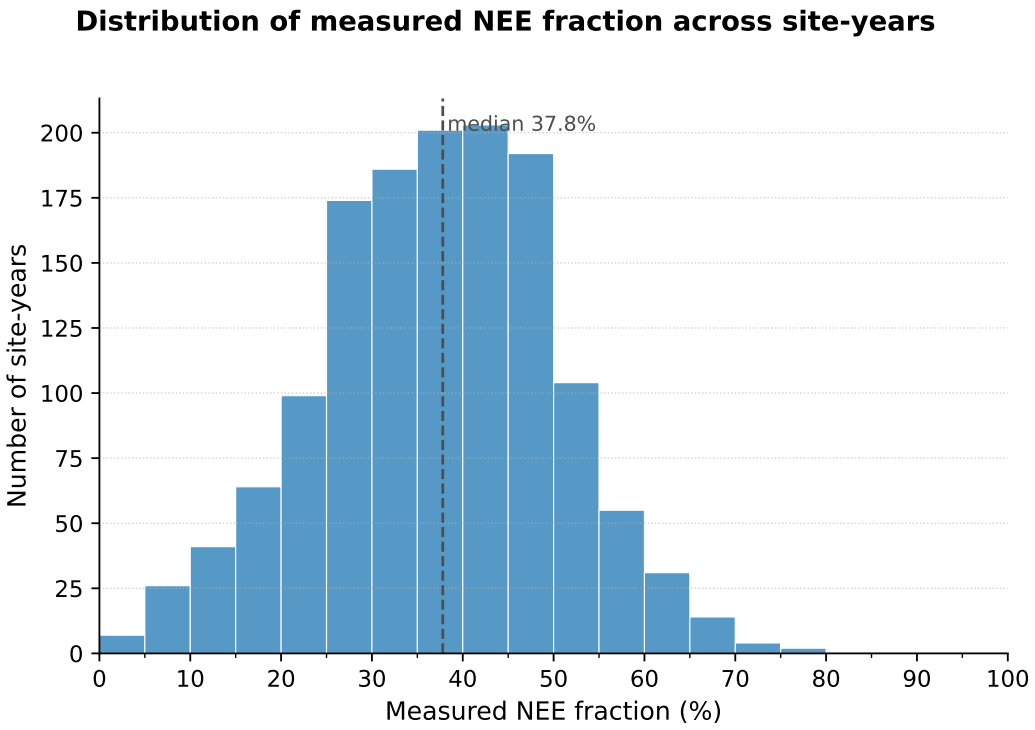


Figure 9: Histogram of data coverage per site-year for the NEE signal. The data coverage is calculated as the percentage of non-missing data points per site-year.

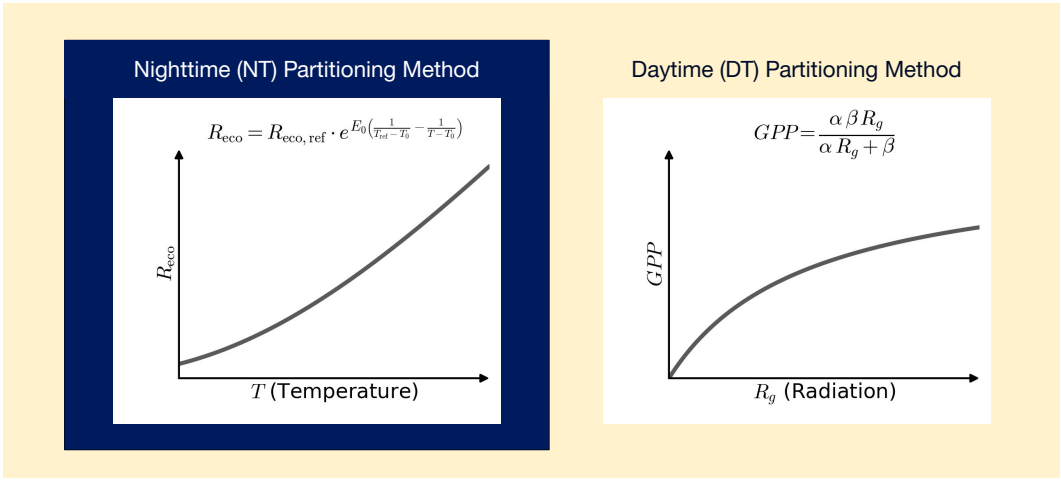


Figure 10: Response functions underlying the parametric partitioning methods: the nighttime method uses the temperature-driven respiration model (left) and the daytime method uses the light-response curve (right) in addition to the same respiration model.

Nighttime-based partitioning The nighttime method is the oldest of the implemented partitioning methods (Reichstein et al., 2005b). It is mostly based on the assumption that at night $\mathrm{GPP} = 0$, implying that all observed NEE at night corresponds to noisy respiration, $\mathrm{R_{eco}} \approx \mathrm{NEE}$. Furthermore, they assume an Arrhenius-type equation (Arrhenius, 1889; Van't Hoff et al., 1899; Lloyd & Taylor, 1994):

Table 5: Temporal aggregation scales used in the evaluation. Let $x_{d,h}$ denote the half-hourly flux value on calendar date d at half-hour slot $h \in \{1, \ldots, 48\}$, with $j(d)$ the day-of-year, $m(d)$ the calendar month, and $y(d)$ the calendar year. $\bar{x}_d = \frac{1}{48} \sum_h x_{d,h}$ denotes the daily mean.

Scale	**Aggregation**	**Description**
Half-Hourly	$x_{d,h}$	Raw half-hourly values. No temporal aggregation applied.
Daily	$\bar{x}_d = \frac{1}{48} \sum_{h=1}^{48} x_{d,h}$	Arithmetic mean over all half-hourly timesteps within each calendar date.
Mean Diurnal Cycle	$\mu_{m,h} = \underset{d:\, m(d)=m}{\text{mean}} \left(x_{d,h}\right)$	Mean flux at each half-hour slot h within each calendar month m, averaged across all days of that month.
Seasonal Cycle	$S_j = \underset{d:\, j(d)=j}{\text{mean}} \left(\bar{x}_d\right)$	Mean of daily means across all years sharing the same day-of-year j. Captures the climatological seasonal pattern without any additional smoothing.
Anomalies	$a_d = \bar{x}_d - S_{j(d)}$	Deviation of each daily mean from the multi-year day-of-year climatology $S_{j(d)}$. Requires > 3 years of data per site.
Inter-Annual Variability	$v_y = \bar{x}_y - \frac{1}{Y} \sum_{y'=1}^{Y} \bar{x}_{y'}, \quad \bar{x}_y = \underset{d:\, y(d)=y}{\text{mean}} \left(\bar{x}_d\right)$	Deviation of each annual mean from the grand mean across all Y years. Requires $Y > 5$ years of continuous data per site.

Table 6: Special evaluation subsets used to assess method performance under challenging conditions.

Subset	**Description**
Rain pulses	A $\pm$2-day window centred on each identified rain pulse event, evaluated at half-hourly resolution. Multiple events within a site-year are treated independently. Targets the rapid ecosystem response to precipitation inputs, particularly the pulse respiration response in water-limited ecosystems.
Unmeasured	Half-hourly time steps where NEE was not directly measured (`NEE_QC` $\neq 0$), i.e. gap-filled periods. Assesses method performance under conditions where no direct observational constraint is available.
Peak temperature	Half-hourly time steps where air temperature exceeds the maximum nighttime temperature recorded in the respective site-year. Targets periods of anomalous heat stress that are not represented in the nighttime temperature envelope used by traditional respiration models.

$$\mathrm{R_{eco}} = R_{\mathrm{eco,ref}} \cdot e^{E_0 \left(\frac{1}{T_{\mathrm{A}}^{\mathrm{ref}} - T_0} - \frac{1}{T_{\mathrm{A}} - T_0} \right)}. \tag{7}$$

The temperature sensitivity E_0 and the reference respiration $R_{\mathrm{eco,ref}}$ are both fitted with moving windows. E_0 is fitted on a 15-day moving window with 5-day steps, with some safeguards for too small temperature ranges, sensible parameter values, and too high error rates. Subsequently, $R_{\mathrm{eco,ref}}$ is fitted with 4-day windows and then linearly interpolated to get smoothly varying values. This model is used to predict $\mathrm{R_{eco}}$ during the day and recover GPP as its residuals following $\mathrm{GPP} = \mathrm{R_{eco}} - \mathrm{NEE}$. While $\mathrm{R_{eco}}$ mostly captures the slow-varying dynamics and temperature dependence, GPP is left with all other variability, including noise. It is hence also not constrained to be non-negative. For this reason, until now nighttime GPP is the preferred product if you want to avoid smoothed-out GPP dynamics, but it comes at the price of relying on sparse and uncertain nighttime data and the extrapolation of $\mathrm{R_{eco}}$ into the daytime conditions.

Table 7: Performance metrics used in the evaluation. Let y_i denote the ground-truth flux and $\hat{y}_i$ the predicted flux at time step i, with $\bar{y}$ and $\bar{\hat{y}}$ the respective sample means.

Metric	Definition	Optimal	Range
Bias	$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$	0	$(-\infty, \infty)$
RMSE	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$	0	$[0, \infty)$
cRMSE	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(\hat{y}_i - \bar{\hat{y}}) - (y_i - \bar{y})\right]^2}$	0	$[0, \infty)$
Correlation	$\left(\frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}\right)^2$	1	$[0, 1]$
NSE	$1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$	1	$(-\infty, 1]$

Table 8: Overview of flux partitioning methods evaluated in this benchmark. Engineered features are described in Appendix B. NT: nighttime partitioning; DT: daytime partitioning; NN: neural network; FE: feature engineering; KG: knowledge-guided; PT: pretraining; USTAR: friction velocity threshold.

Method	Description	Reference	GPP inputs	RECO inputs
NT	Nighttime-based partitioning. Fits an exponential temperature sensitivity of respiration to nighttime NEE; GPP obtained by subtraction.	Reichstein et al. (2005b)	Derived by subtraction	NEE (night), TA, USTAR
DT	Daytime-based partitioning. Fits a light-response curve to daytime NEE to estimate GPP and RECO simultaneously.	Lasslop et al. (2010)	SW_IN, TA, VPD	TA, VPD
DT_{Keenan}	Modified daytime partitioning. As DT but nighttime RECO is estimated at a reference temperature rather than extrapolated.	Keenan et al. (2019)	SW_IN, TA, VPD	TA (ref. temp.)
NN_{Base}	LUE-based neural network baseline following Tramontana et al. (2020). Separate sub-networks for GPP (LUE formulation) and RECO, trained with MSE loss.	Tramontana et al. (2020)	SW_IN, SW_IN_POT, VPD, TA, SWC, WS, SW_IN_POT_diff, SW_IN_POT_avg, SW_IN_POT_avg_diff, GPP_prox	TA, TS, SWC, WS, doy_sin, doy_cos, NEE_nt_avg
NN_{L1}	As NN_{Base} with L1 loss replacing MSE. Serves as baseline for extensions below.	—	Same as NN_{Base}	Same as NN_{Base}
NN_{FE}	As NN_{L1} with an extended set of engineered features targeting light-use efficiency proxies, nighttime dynamics, and soil moisture memory.	—	TA, RH, SW_IN, SW_IN_POT, SW_IN_POT_diff, NEE_SWin_lowlight_lm_slope, NEE_LUE_at_max_SW, NEE_light_night_diff, SWC_capa, WS, HR_sin, HR_cos	NEE_night_movmean, TA, TS, TS_night_movmean, TA_night_movmean, SWC_capa, harmonics, WS, HR_sin, HR_cos
NN_{KG}	As NN_{L1} with knowledge-guided regularisation: RECO predictions are penalised toward the NT partitioning product during training.	—	Same as NN_{L1}	Same as NN_{L1}
NN_{PT}	As NN_{L1} with a pretraining stage before joint fine-tuning.	—	Same as NN_{L1}	Same as NN_{L1}
NN_{Night}	Minimal-structure network that constrains GPP to zero at night. No LUE formulation; daytime GPP predicted directly.	Zhan et al. (2022)	TA, RH, SWC, SW_IN, WS, EF_avg	TA, SWC, WS, NEE_nt_avg, EF_avg
DML	Double machine learning framework. Estimates GPP via a causal, partially linear model using gradient boosting residualisation.	—	VPD, TA, CWD, doy_sin, doy_cos (causal variables); RECO: TA, CWD, WS, doy_sin, doy_cos, NEE_nt_avg	

Daytime-based partitioning Instead, the daytime method assumes an additional rectangular hyperbola light-response curve for GPP on top of eq. (7):

$$\mathrm{NEE} = \frac{\alpha\beta \mathrm{SW}_{\mathrm{IN}}}{\alpha \mathrm{SW}_{\mathrm{IN}} + \beta} + R_{\mathrm{eco,ref}} e^{E_0\left(\frac{1}{T_{\mathrm{A}}^{\mathrm{ref}} - T_0} - \frac{1}{T_{\mathrm{A}} - T_0}\right)}. \tag{8}$$

The parameter α is the initial slope and corresponds to the initial light-use efficiency, and β corresponds to uptake at light saturation. The sensitivity parameter E_0 is again estimated from nighttime data in a first step. After this, the remaining $R_{\mathrm{eco,ref}}$, α and β are estimated on daytime data on short windows. To make the uptake adapt to extreme lack of water, β has a reduction mechanism based on

the threshold $\mathrm{VPD}_0 = 10\,\mathrm{hPa}$ (Körner, 1995):

$$\beta = \begin{cases} \beta_0\, e^{-k(\mathrm{VPD}-\mathrm{VPD}_0)}, & \mathrm{VPD} > \mathrm{VPD}_0, \\ \beta_0, & \mathrm{VPD} \le \mathrm{VPD}_0, \end{cases}$$

which also requires to fit $k \geq 0$ on windows. As $R_{\mathrm{eco,ref}}$ is inferred from daytime data, it is not so affected by the turbulence-related biases at night. An advantage of the VPD relationship is that it can capture the afternoon-depressed diurnal cycle, but on the other hand, the fixed functional form with only T_{A}, $\mathrm{SW_{IN}}$ and VPD as drivers is a strong limitation on the effects the resulting fluxes can capture.

Both methods clearly smooth out abrupt changes and introduce structural errors. As both methods rely on largely different data and assumptions, it is usually taken as a sign of a reliable partitioning product when both methods agree.

Modified daytime-based partitioning (Keenan et al., 2019) developed an extension of the daytime method that accounts for the Kok effect (inhibition of respiration under strong light). The Kok effect is hard to capture with methods that constrain the effect of light on respiration with nighttime data alone. This will lead to an overestimation of respiration during the day. A light-inhibition factor is introduced to reduce $\mathrm{R_{eco}}$ during daylight. Neglecting it is shown to introduce substantial bias into both fluxes.

We run all parametric methods using the REddyProc implementation (Wutzler et al., 2018).

B.2 Machine Learning Baselines

The ML models are a set of extensions based on previously published works.

B.2.1 Neural Models

Minimum Structure Model $\mathrm{NN_{night}}$ The $\mathrm{NN_{night}}$ corresponds to the `no-sif` architecture from Zhan et al. (2022), which imposes minimal structure: no explicit form for GPP, only the constraint that it vanishes at night. Specifically, it implements

$$\mathrm{NEE} = -\mathrm{GPP}(X_{\mathrm{GPP}}) \cdot \mathbb{1}_{\mathrm{DAY}}(t) + \mathrm{R_{eco}}(X_{\mathrm{R_{eco}}}), \tag{9}$$

with one neural network implementing GPP and a second one $\mathrm{R_{eco}}$.

Light-use Efficiency Model Tramontana et al. (2020) popularized an implementation of the light-use efficiency model, namely:

$$\mathrm{NEE} = -\mathrm{LUE}(X_{\mathrm{LUE}}) \cdot \mathrm{SW_{IN}} + \mathrm{R_{eco}}(X_{\mathrm{R_{eco}}}) \tag{10}$$

where a neural network is used to implement LUE and $\mathrm{R_{eco}}$. As the rectangular hyperbola eq. (8) that is assumed in the daytime method suggests, GPP is not really linear in $\mathrm{SW_{IN}}$. To provide it with the necessary flexibility, $\mathrm{SW_{IN}}$ is generally included among the inputs to the LUE.

Architecture and training All neural baselines share common backbones and training routines. Each consists of two multilayer perceptrons, one for GPP (or LUE) and one for $\mathrm{R_{eco}}$. Each has two hidden layers with 64 units, `ReLU` activations, and a final softplus nonlinearity to keep the output positive. All input features for neural networks are standardized. Only $\mathrm{SW_{IN}}$ as a factor of GPP in the LUE model and the target NEE are normalized to $[0, 1]$. We fit each model per site-year. We optimize all models with Adam with an initial learning rate of 10^{-3} (Kingma & Ba, 2017). The $\mathrm{NN_{base}}$ model minimizes an mean squares error (MSE) objective. All other models minimize the mean absolute error (MAE), i.e., the L_1 loss, matching the Laplacian noise model. In order to make sure that we encompass a range of learning rates and converge properly, we use early stopping with a reduce-on-plateau scheduler. 30% of the data are held out for validation. Every time the validation loss does not reduce during 7 epochs, we reset to the best registered model of the learning rate level and reduce the learning rate by a factor of 0.3. After the 7 reduction steps, we reset to the best model on the validation set seen during training. We limit the training to a maximum of 400 epochs. For each site-year, we train an ensemble of 10 models, each with a different random train/validation split and initialization. We average their predictions to obtain the final partitioning.

Note that we used the general network architecture proposed by Zhan et al. (2022), though we applied a more elaborate training routine with scheduler. Tramontana et al. (2020) optimized their much smaller network (a single hidden layer with 10-15 hidden units) using Levenberg-Marquardt, which was sensible for their deep learning framework in MATLAB. Modern cutting-edge deep learning frameworks rely on efficient optimizers such as Adam and overparameterization, which gives more flexibility.

B.2.2 EXTENSIONS

We implement various extensions based on ML techniques such as feature engineering, transfer learning or regularization. These methods are designed to tackle very specific issues in flux partitioning.

$\mathrm{NN_{KG}}$ This extension uses targets from the nighttime method as a regularizer for both fluxes. The regularizer is implemented through the correlation coefficient. In this way, the method encourages covariation but not exact correspondence in magnitude. Concretely, for each component $f \in \{\mathrm{GPP}, \mathrm{R_{eco}}\}$ we take the model prediction $\hat{f}$ and the NT estimate f^{NT} over the site-year and penalize

$$\mathcal{L}_f = 1 - \rho\left(\hat{f},\, f^{\mathrm{NT}}\right), \tag{11}$$

where ρ denotes the Pearson correlation.

To make the regularization adaptive rather than requiring a fixed manual trade-off (which may need to change between site-years), we treat each guidance signal as a noisy target and learn its weight using the homoscedastic uncertainty scheme of Kendall et al. (2017). In this framework, each objective $\mathcal{L}_t$ for a task t over a set of tasks T is connected to a learnable task-dependent uncertainty σ_t. We illustrate the idea with a Gaussian likelihood

$$p(\{y_t\} \mid \{\mu_t\}) = \prod_{t \in T} \mathcal{N}(y_t; \mu_t, \sigma_t^2)$$

with task prediction μ_t, and standard deviation σ_t. The negative $\log$-likelihood lets us derive for MSE $\mathcal{L}_t$

$$\mathcal{L}(\{\mu_t\}, \{\sigma_t\}) = \sum_{t \in T} \frac{1}{2\sigma_t^2} \mathcal{L}_t(\mu_t) + \log \sigma_t, \tag{12}$$

in which the $1/2\sigma_t^2$ term down-weights uncertain objectives and the $\log \sigma_t$ term prevents the trivial solution $\sigma_t \to \infty$. We keep the NEE reconstruction $\mathcal{L}_{\mathrm{NEE}}$ as the primary objective at unit weight and apply eq. (12) only to the two correlation regularizers, giving the full training objective

$$\mathcal{L}\left(\{\hat{\mathrm{GPP}}, \hat{\mathrm{R_{eco}}}\}, \{\sigma_{\mathrm{GPP}}, \sigma_{\mathrm{R_{eco}}}\}\right) = \mathcal{L}_{\mathrm{NEE}}(\hat{\mathrm{GPP}}, \hat{\mathrm{R_{eco}}}) \tag{13}$$

$$+ \sum_{f \in \{\mathrm{GPP}, \mathrm{R_{eco}}\}} \left[\frac{1}{2\sigma_f^2} \big(1 - \rho(\hat{f}, f^{\mathrm{NT}})\big) + \log \sigma_f \right], \tag{14}$$

where each σ_f is parametrized through $\log \sigma_f$ for positivity and optimized jointly with the network weights[4]. In particular, the weighting is adaptive. The guidance is pulled in strongly where the model can align with the NT pattern and relaxed where it hinders fitting the loss.

$\mathrm{R_{eco}}$ is generally less constrained by data, as it is the flux with lower magnitude. Sometimes ML models invert the daily cycle. This is not physical but rather an artifact of poor training. With this

[4]Even though eq. (12) illustrates this framework for Gaussian objectives, the correlation loss can also be derived from a likelihood. We write $\tilde{f}$ for the mean-centered and unit-normalized version of a vector f over the site-year. This yields $\|\tilde{\hat{f}} - \tilde{f}^{\mathrm{NT}}\|^2 = 2\big(1 - \rho(\hat{f}, f^{\mathrm{NT}})\big)$, so $\mathcal{L}_f$ is (half) the squared error of the *whitened* target under $\mathcal{N}(\tilde{f}^{\mathrm{NT}}; \tilde{\hat{f}}, \sigma_f^2 I)$. This recovers exactly the $1/2\sigma_f^2$ weight and $\log \sigma_f$ normalizer of eq. (12). Because ρ is invariant to affine transformations of $\hat{f}$, this term constrains only the temporal *shape* of the fluxes toward the NT product. Their magnitude is fixed by the NEE data term.

regularization, we implicitly encourage $\mathrm{R_{eco}}$ to behave similarly to the parametric formulation of $\mathrm{R_{eco}}$, which, by design, shows a diurnal cycle that peaks around noon, when temperatures are highest. In this way, this method aims to hedge against little data and low variation at night that could lead to a poor functional behavior of $\mathrm{R_{eco}}$.

$\mathrm{NN_{PT}}$ We pretrain on the two years before and the two years after (four in total). This is a form of transfer learning: the goal is to capture the general seasonal dynamics through the more abundant data over multiple years and only adapt it to the single year. Like the regularization method, this can mitigate the effects of insufficient data, but it acts through initialization rather than through an added loss term. One advantage of pretraining is that the actual loss does not need to compete with other objectives as in the case of regularization, which pulls toward the smoother parametric behavior of NT. At the same time, it is very easy to just overwrite the pretraining during fine-tuning, rendering it ineffective. We attempt to prevent this by setting a lower learning rate during fine-tuning.

$\mathrm{NN_{FE}}$ This method uses various designed features that carry relevant memory and dynamics: light-use-efficiency proxies, nighttime dynamics, and soil-moisture memory. This should allow even a static model to adapt rapidly to dynamic effects such as rain pulses. Specifically, we add

- *Light-response proxies.* `NEE_SWin_lowlight_lm_slope` is the slope of a linear fit of NEE against $\mathrm{SW_{IN}}$ in the low-light regime, i.e., the initial slope of the light-response curve and thus a proxy for the initial LUE. `NEE_LUE_at_max_SW` is the local slope $\partial\mathrm{NEE}/\partial\mathrm{SW_{IN}}$ evaluated near peak radiation, which is a proxy for the light-saturated efficiency. `NEE_light_night_diff` is the difference between the smoothed daytime and nighttime NEE, which approximates the net daytime uptake ($\approx$ GPP), since nighttime NEE $\approx \mathrm{R_{eco}}$.
- *Nighttime dynamics.* `NEE_night_movmean`, `TA_night_movmean`, `TS_night_movmean` are rolling (median-filtered) summaries of nighttime NEE, air temperature and soil temperature that carry the recent nighttime respiration and temperature information as a memory signal.
- *Soil-moisture memory.* `SWC_capa_`c, $c \in \{1, 5, 25, 125, 625\}$, is a leaky-bucket surface-moisture state driven by precipitation and drained by evapotranspiration (`LE`), where the capacity c sets the memory timescale. The family spans fast to slow drainage so the model can respond both to rain pulses and to prolonged drought.

B.2.3 DML-HM

Finally, we include a DML approach (Cohrs et al., 2024), which estimates GPP through a causal, partially linear model: Nuisance functions for the outcome and treatment components, and the heterogeneous effect light-use efficiency (LUE), are each fit with gradient-boosted regression trees Friedman (2001). Finally, the respiration is learned based on the difference between NEE and the estimated GPP with a NN and a new set of variables $X_{\mathrm{R_{eco}}}$ (see table 8).

Problem adaptation As mentioned in the context of the LUE, GPP is not linear in $\mathrm{SW_{IN}}$, which prompted us to introduce $\mathrm{SW_{IN}}$ as a driver of LUE. The same problem prohibits simply treating the problem as a partially linear regression problem with heterogeneous treatment effect (the function LUE). Instead, Cohrs et al. (2024) proposes to transform the problem in the following way

$$\mathrm{NEE} = -\mathrm{LUE}(X_{\mathrm{mod}}) \cdot q(\mathrm{SW_{IN}}) + \mathrm{R_{eco}}(X_{\mathrm{mod}}, X_{\mathrm{ctrl}}), \tag{15}$$

where X_{mod} and X_{ctrl} are sets of meteorological drivers and q transforms the incoming radiation to allow for more flexible light-response curves, leading to a potentially non-linear light-use efficiency model. Here, $\mathrm{R_{eco}}(\cdot)$ and $\mathrm{LUE}(\cdot)$ take the roles of the outcome nuisance $g(\cdot)$ and the effect function $\theta(\cdot)$ of the equivalent partially linear model, respectively. We follow the original implementation and fit a rectangular hyperbola over a moving window of 15 days to obtain $q(\mathrm{SW_{IN}})$. We fit a final estimator for $\mathrm{R_{eco}}$ on the residuals.

Causal graph of the LUE **model** The causal graph assumed for the LUE model is shown in fig. 11 (Cohrs et al., 2024). Consequently, this method estimates GPP as the direct effect of light on NEE, discounting the indirect effects through temperature T_{A}, vapor pressure deficit VPD, and

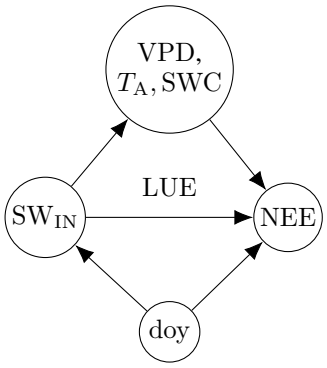


Figure 11: Assumed causal graph for the estimation with the causal hybrid modeling approach in flux partitioning (Cohrs et al., 2024).

soil water content SWC, which we allocate to R_{eco}. Additionally, we control for day of the year doy, which represents seasonality. Note that these variables still enter as modifiers on the effect of light on NEE, affecting GPP.

C Additional Results

C.1 Best-method Share

Table 9 reports, for each method, the share of site-years in which it achieves the best score for a given metric, complementing the aggregate view given in the main text.

Table 9: Best-method share, defined as the percentage of site-years for which each method achieves the best score for a given metric (NSE, maximized and Bias, minimized in absolute value), reported separately for NEE, GPP, and RECO, and for QUINCY (Q) and CABLE-POP (C-P). In case of a tie, both methods are counted, which can lead to a sum of percentages exceeding 100%.

	NEE				GPP				RECO			
	NSE		Bias		NSE		Bias		NSE		Bias	
	Q	C-P	Q	C-P	Q	C-P	Q	C-P	Q	C-P	Q	C-P
NT	2.3%	1.5%	16.7%	17.3%	1.4%	0.6%	**23.1%**	6.6%	17.4%	6.1%	**25.0%**	12.4%
DT	0.1%	0.1%	4.1%	1.9%	1.4%	0.4%	9.3%	5.9%	0.4%	0.2%	4.8%	2.6%
TK	1.4%	0.4%	9.7%	6.1%	1.4%	0.4%	9.3%	5.9%	0.4%	0.2%	7.8%	4.6%
DML	0.0%	0.1%	0.7%	0.1%	0.4%	0.3%	6.9%	6.8%	0.1%	0.1%	2.3%	0.6%
NN_{base}	3.0%	2.2%	5.3%	3.4%	2.6%	3.7%	4.9%	8.3%	0.4%	2.1%	1.9%	3.5%
NN_{KG}	16.4%	13.1%	17.5%	21.3%	**51.4%**	10.8%	21.1%	17.9%	**53.7%**	29.2%	24.8%	22.0%
NN_{L1}	**47.8%**	22.1%	9.3%	8.2%	21.5%	31.8%	3.3%	15.5%	6.6%	18.5%	3.1%	10.8%
NN_{FE}	1.7%	3.6%	7.6%	8.3%	2.5%	2.7%	13.1%	9.6%	7.2%	3.7%	12.0%	12.7%
NN_{PT}	42.1%	**58.3%**	**19.2%**	**26.0%**	29.7%	**54.0%**	6.9%	**22.8%**	15.0%	**41.6%**	6.9%	**23.2%**
NN_{night}	2.6%	9.3%	13.7%	12.1%	2.1%	5.9%	12.6%	9.2%	1.5%	2.1%	12.7%	9.4%

C.2 Performance spread of partitioning methods across different scales and events

Figures 12 to 15 show the full per-site-year NSE distributions behind the medians reported in the main text, separately for GPP and R_{eco} and for each simulator, across temporal scales and stress-event conditions.

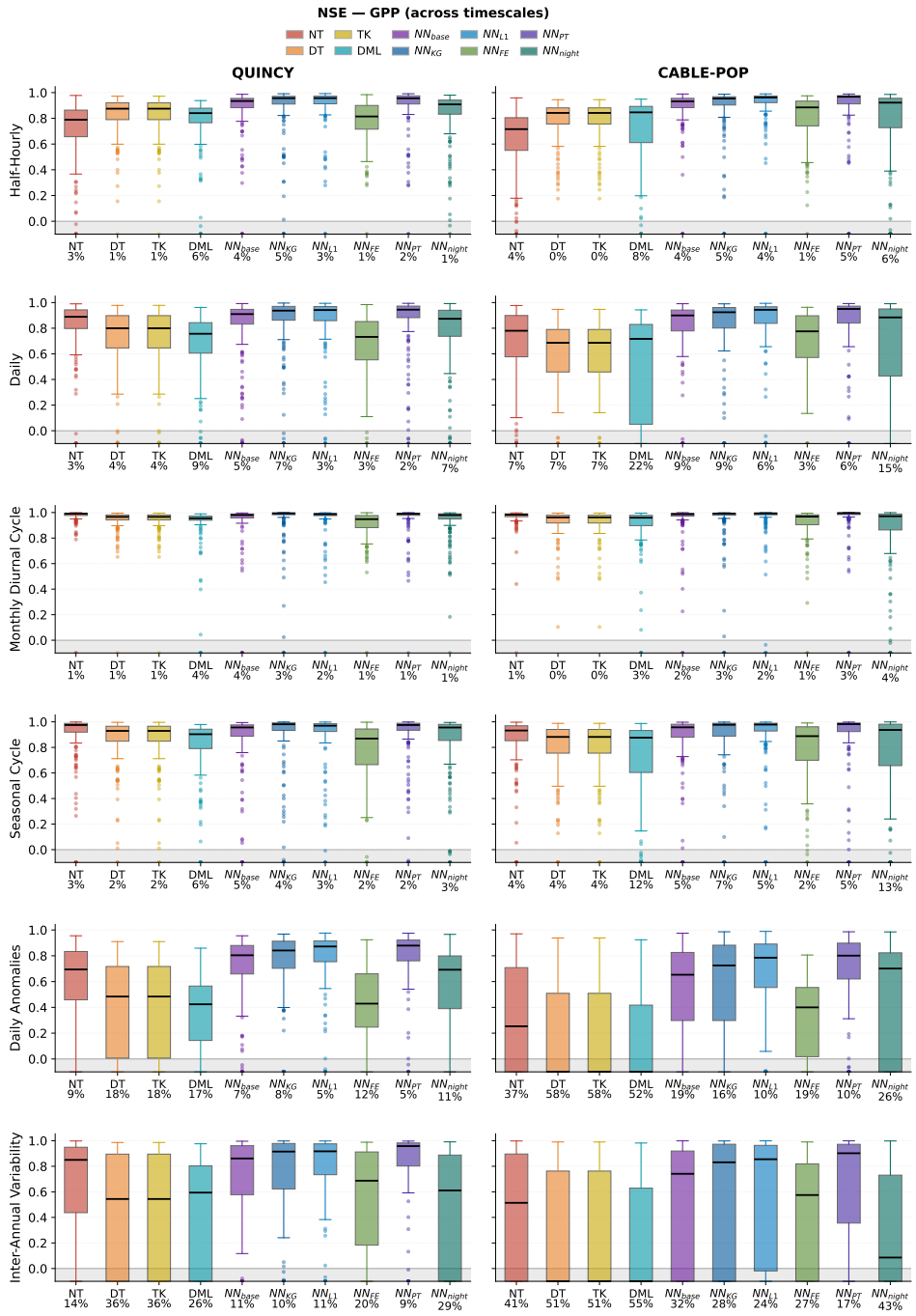


Figure 12: Distribution of GPP partitioning NSE performance across temporal scales. Boxplots show per-site-year NSE distributions for each method, separately for QUINCY (left) and Cable-Pop (right), across half-hourly, daily, monthly diurnal cycle, seasonal cycle, inter-annual anomalies, and inter-annual variability. Axes are clipped at -0.1 (shaded band) to keep the bulk of the distribution readable. Site-years scoring below this threshold are not shown and are excluded from the visible box/whisker geometry.

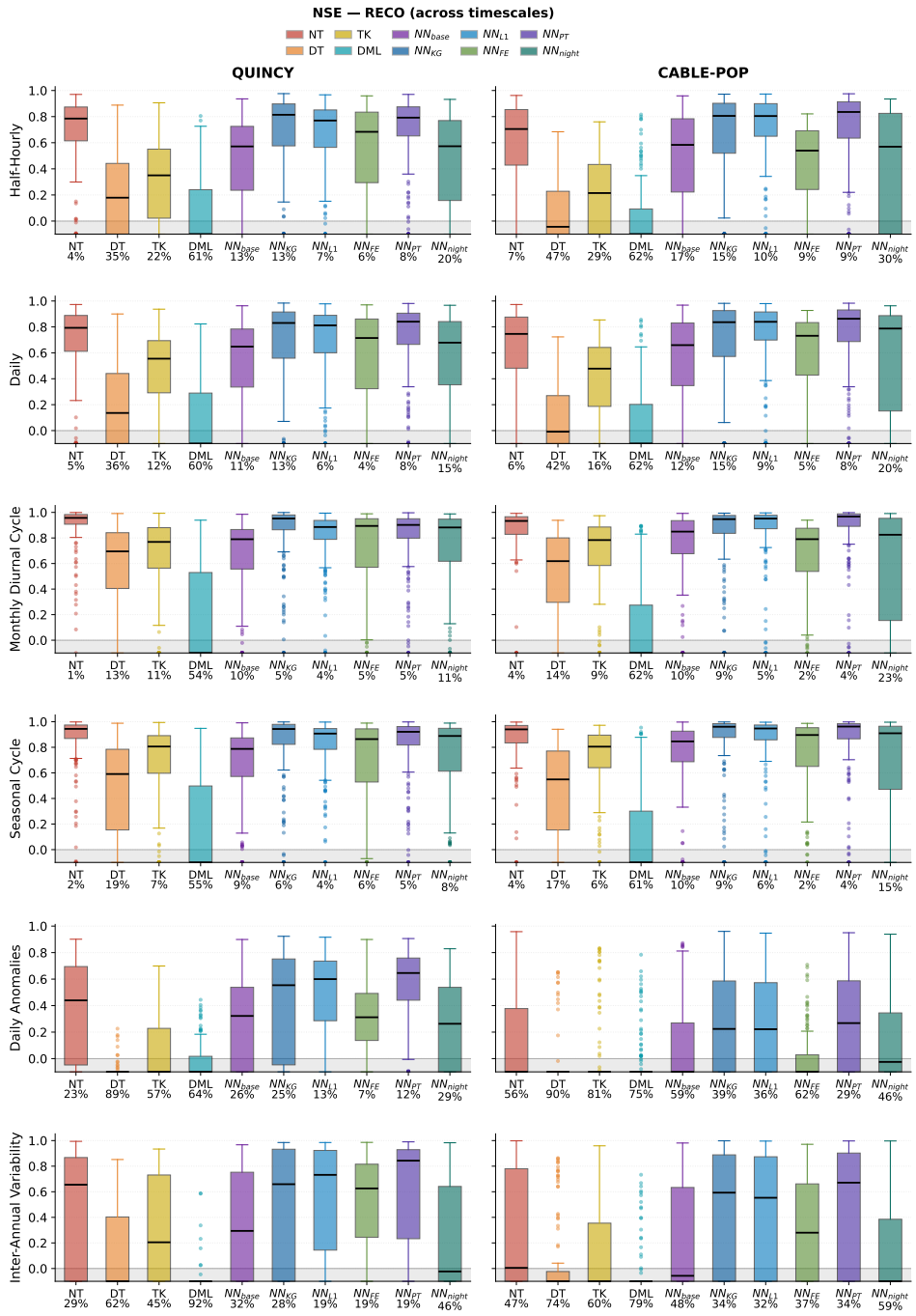


Figure 13: Distribution of partitioning NSE performance across temporal scales. Boxplots show per-site-year NSE distributions for each method, separately for QUINCY (left) and Cable-Pop (right), across half-hourly, daily, monthly diurnal cycle, seasonal cycle, inter-annual anomalies, and inter-annual variability. Axes are clipped at -0.1 (shaded band) to keep the bulk of the distribution readable. Site-years scoring below this threshold are not shown and are excluded from the visible box/whisker geometry.

C.3 Stress-regime analysis: identifying failure modes

This section collects the per-regime figures behind the failure-mode analysis of section 4.2, extending the main-text results across all IGBP classes and both simulators: coverage-by-IGBP breakdowns

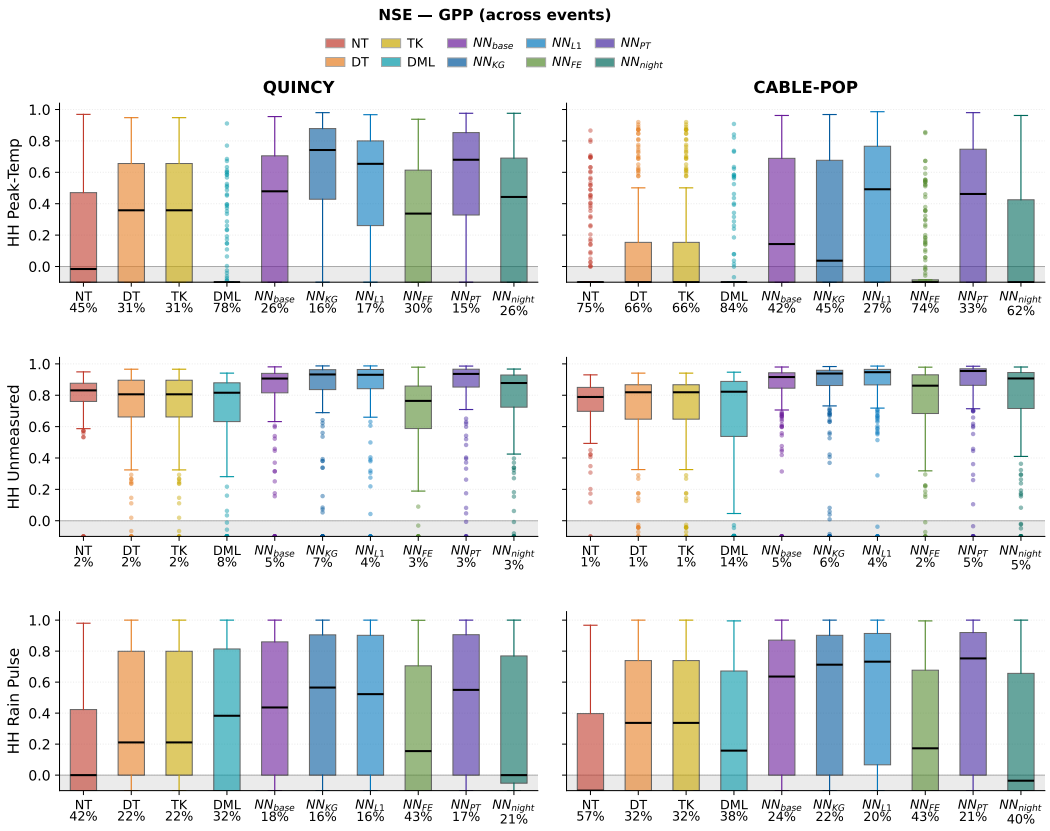


Figure 14: Distribution of GPP partitioning NSE across challenging conditions. Boxplots show per-site-year NSE distributions for each method, separately for QUINCY (left) and Cable-Pop (right), across peak-temperature days, unmeasured/gap-filled timesteps, and rain-pulse events. Axes are clipped at -0.1 (shaded band) to keep the bulk of the distribution readable. Site-years scoring below this threshold are not shown and are excluded from the visible box/whisker geometry.

(figs. 16 to 18), performance across climate space (figs. 19 and 20), the diurnal warm/cool-day error for CABLE-POP (fig. 21), and the rain-pulse composite error together with two single-event examples (figs. 22 and 23). Each is discussed at the corresponding point in section 4.2.

C.4 Transfer to measured data

A major question is, finally, whether the findings on the benchmark carry any meaning for partitioning on real-world data from flux towers.

Summary We analyze how well the behavior of the partitioning methods on the two models transfers to behavior on measured data from the fluxtowers.

- Climate-space plot establishes the general principle (broad regime patterns transfer) and shows the warm-zone dark cluster is reproduced.
- Binning by data coverage shows that the benchmark matches the measured spread at decent coverage, but the match degrades at the sparsest coverage.
- Peak-T diurnal plot shows that the benchmark captures real-data disagreement structure.
- The pulse plot shows that the benchmark exhibits transfer failures for event-specific dynamics, identifying a specific gap.

Disagreement between methods is computed across a fixed set of five (NT, DT, DML, NN_{L1}, $\mathrm{NN}_{\mathrm{KG}}$, $\mathrm{NN}_{\mathrm{FE}}$). For the inter-method spread, all three fluxes (NEE, GPP, $\mathrm{R}_{\mathrm{eco}}$) are normalized

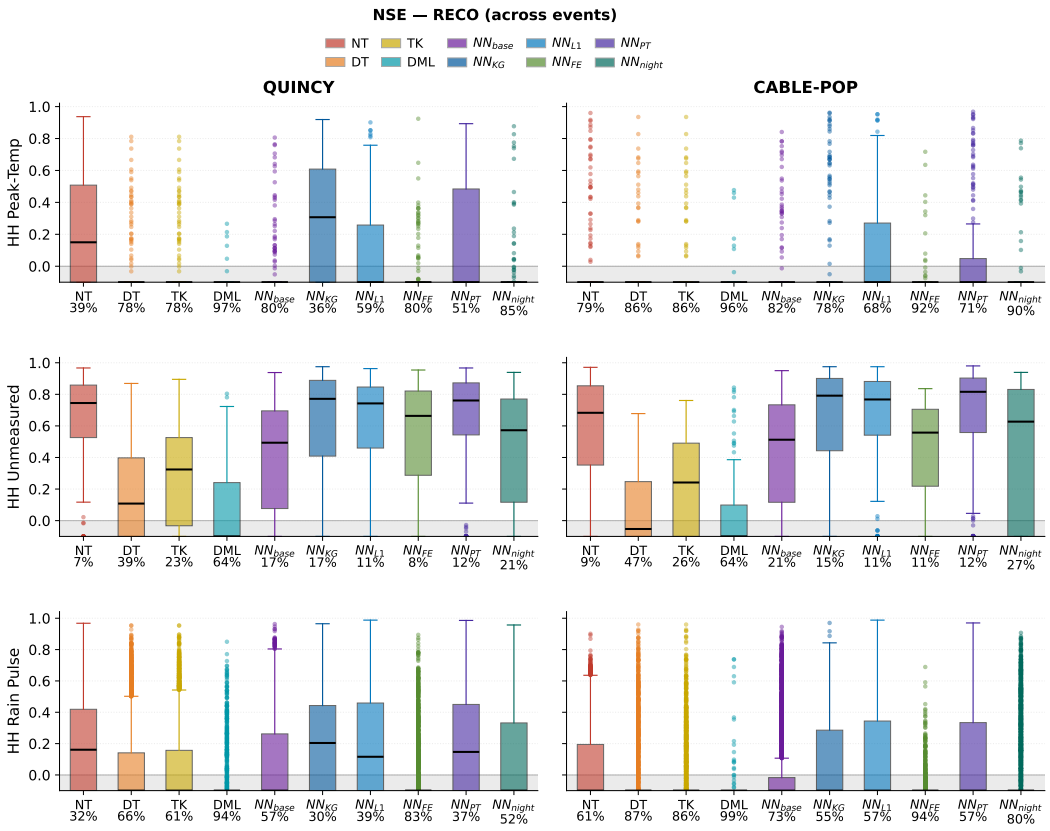


Figure 15: Distribution of partitioning NSE performance across challenging event conditions. Boxplots show per-site-year NSE distributions for each method, separately for QUINCY (left) and Cable-Pop (right), across peak-temperature days, unmeasured/gap-filled timesteps, and rain-pulse events. Axes are clipped at -0.1 (shaded band) to keep the bulk of the distribution readable. Site-years scoring below this threshold are not shown and are excluded from the visible box/whisker geometry.

by the standard deviation of NEE in the ground truth (simulated models) or measurement record (FLUXNET), since ground-truth GPP and R_{eco} are not directly observable for real eddy-covariance data. Consequently, normalized spread values are comparable across climate cells, coverage bins, and data sources within a given flux, but not directly comparable in magnitude across fluxes.

Climate zones We first look at the spread of the methods across climate zones, shown in Figure 24. The plot shows that the spatial pattern of disagreement in the climate space transfers from the simulated to the measured data. The fields show similar gradients from high agreement for temperate-mesic conditions towards higher disagreement in the warmer and wetter climate zones. Reassuringly, the high-disagreement zones coincide roughly with where partitioning accuracy degraded on the benchmark (fig. 19): the inter-method spread, the only signal available on real data, tracks the true error we can measure only in simulation.

Across IGBP Second, we look at the spread of the methods across different IGBP classes which is shown in figure 25. For IGBP classes with sufficient amount of samples over the coverage bins the spread shows a consistent decline with increasing data coverage over both measured and simulated data. From the second bin (15-30%) upward the curves track each other closely in shape and magnitude. Hence, the benchmark reproduces how data availability drives method disagreement across biome types. The match weakens only in the sparsest bin (0-15%) with no consistent direction: the simulators overestimate the measured spread in some classes (DBF) and underestimate it in

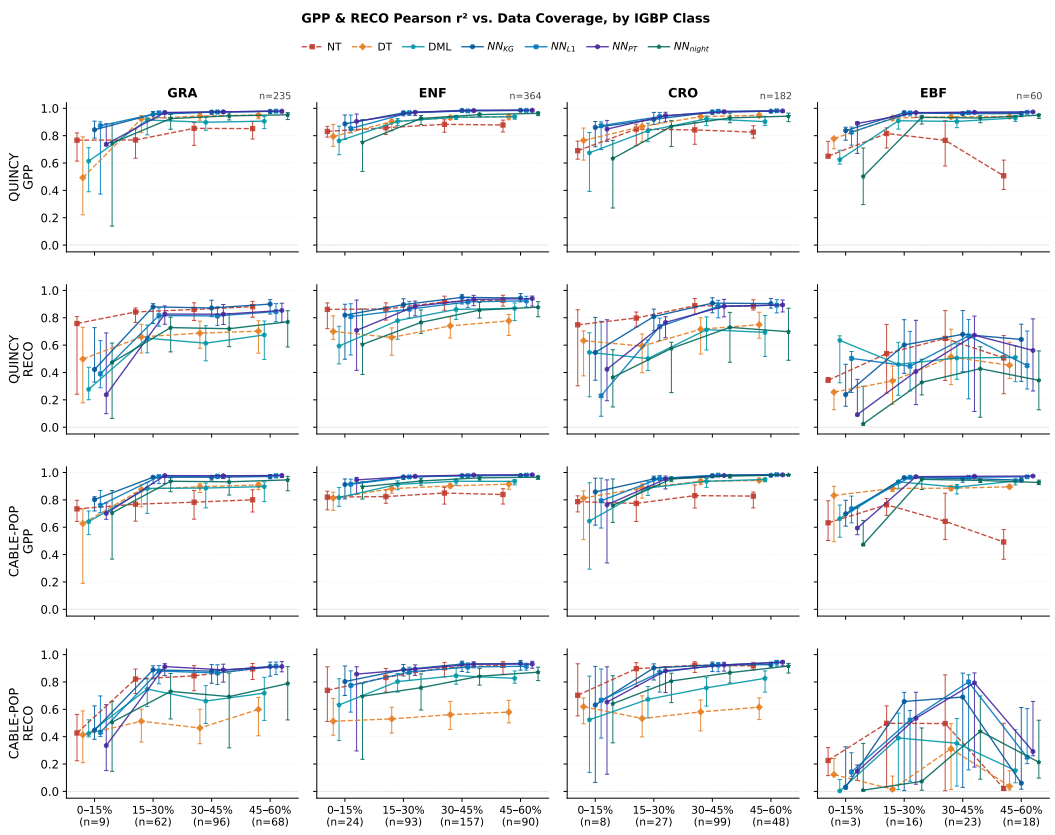


Figure 16: Partitioning performance by data coverage and IGBP land-cover class. Squared Pearson correlation (r^2) at half-hourly resolution for a subset of methods (see section C.3 for selection rationale), shown separately for fluxes and simulators, aggregated within coverage bins up to 60% for four IGBP classes: grassland (GRA), evergreen needleleaf forest (ENF), cropland (CRO), and evergreen broadleaf forest (EBF). Error bars show interquartile range (IQR) across site-years within each bin, with the number of site-years n per bin shown in parentheses and the total n per IGBP class shown at the top. Bins with $n < 10$ should be interpreted with caution.

others (GRA, ENF). There is one notable exception to this pattern: on croplands, the $\mathrm{R_{eco}}$ spread on measured data keeps declining with coverage while both simulators stay roughly flat.

Peak temperature Next, we examine the spread of the methods across the diurnal cycle separated by cool and warm days which is shown in figure 26. The plot shows for $\mathrm{R_{eco}}$ and GPP the spread within the simulations very cleanly transfers to the measured data showing the diurnal structure peaking around midday and elevated for warm days. The warm/cool day separation specifically shows the benchmark captures the extrapolation regime that is a known failure mode. Only for respiration on hot days is the spread slightly higher at noon for measured data. On measured data, the NEE spread peaks around 6 and slightly falls off in the course of the day while QUINCY and CABLE-POP show slight differences in the spread pattern with their peaks around the early afternoon.

Rain pulses Finally, Figure 27 shows the spread of the methods across rain-pulse events. The plot shows that the spread of the methods across rain-pulse events diverges between measured and simulated data. Both simulators show a lower and less immediate increase in spread for all three fluxes after the SOP than the measured data. This suggests that real-world pulses carry a much cleaner signal for the methods to respond to than simulated data. This clearly indicates that the benchmark requires more realistic pulse complexity in future iterations.

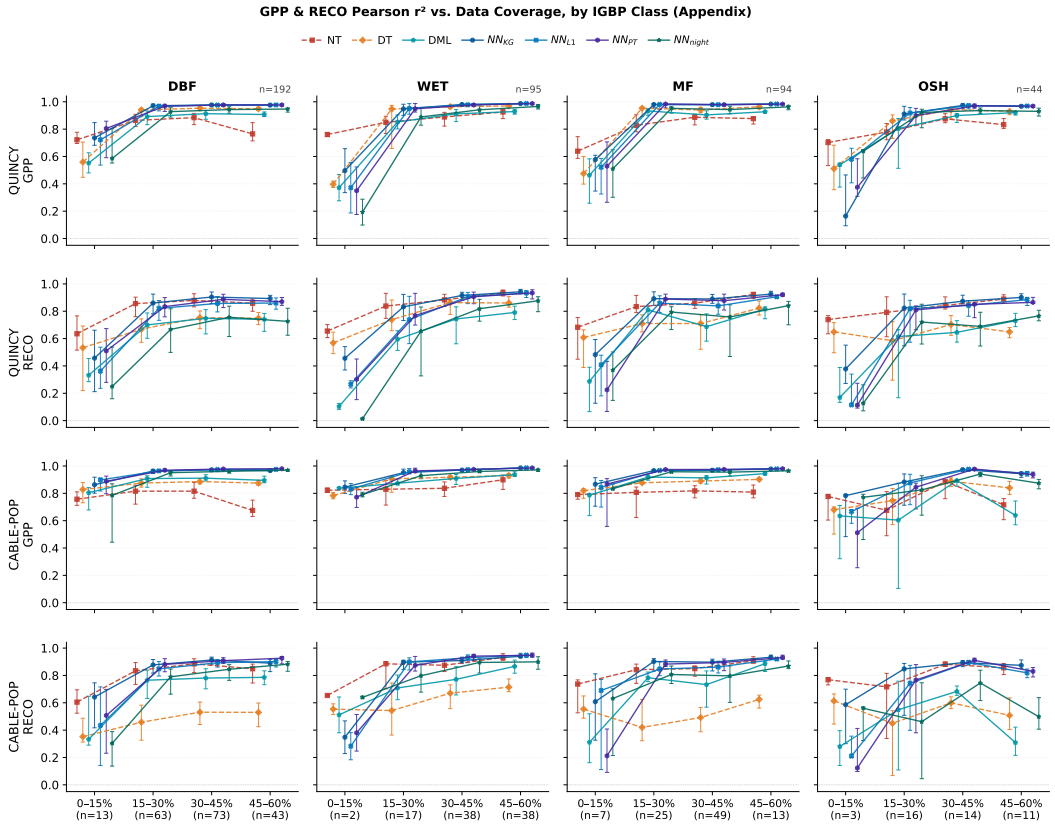


Figure 17: Partitioning performance by data coverage and IGBP land-cover class. Squared Pearson correlation (r^2) at half-hourly resolution for a subset of methods (see section C.3 for selection rationale), shown separately for fluxes and simulators, aggregated within coverage bins up to 60% for IGBP classes: deciduous broadleaf forest (DBF), wetland (WET), mixed forest (MF), and open shrubland (OHS). Error bars show IQR across site-years within each bin, with the number of site-years n per bin shown in parentheses and the total n per IGBP class shown at the top. Bins with $n < 10$ should be interpreted with caution.

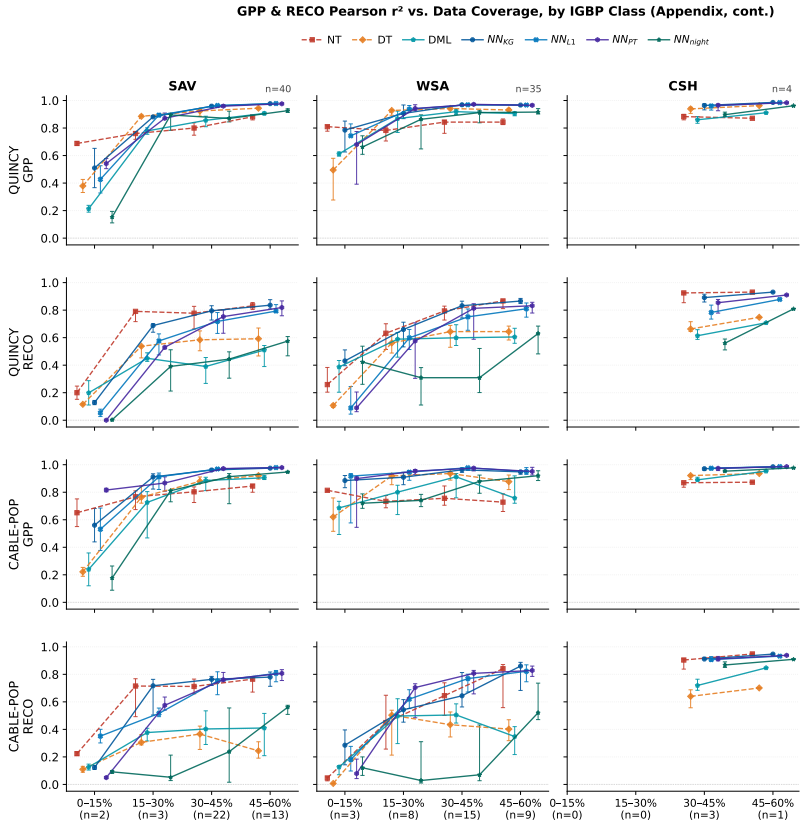


Figure 18: Partitioning performance by data coverage and IGBP land-cover class. Squared Pearson correlation (r^2) at half-hourly resolution for a subset of methods (see section C.3 for selection rationale), shown separately for fluxes and simulators, aggregated within coverage bins up to 60% for IGBP classes: savanna (SAV), woody savanna (WSA), and closed shrubland (CSH). Error bars show IQR across site-years within each bin, with the number of site-years n per bin shown in parentheses and the total n per IGBP class shown at the top. Bins with $n < 10$ should be interpreted with caution.

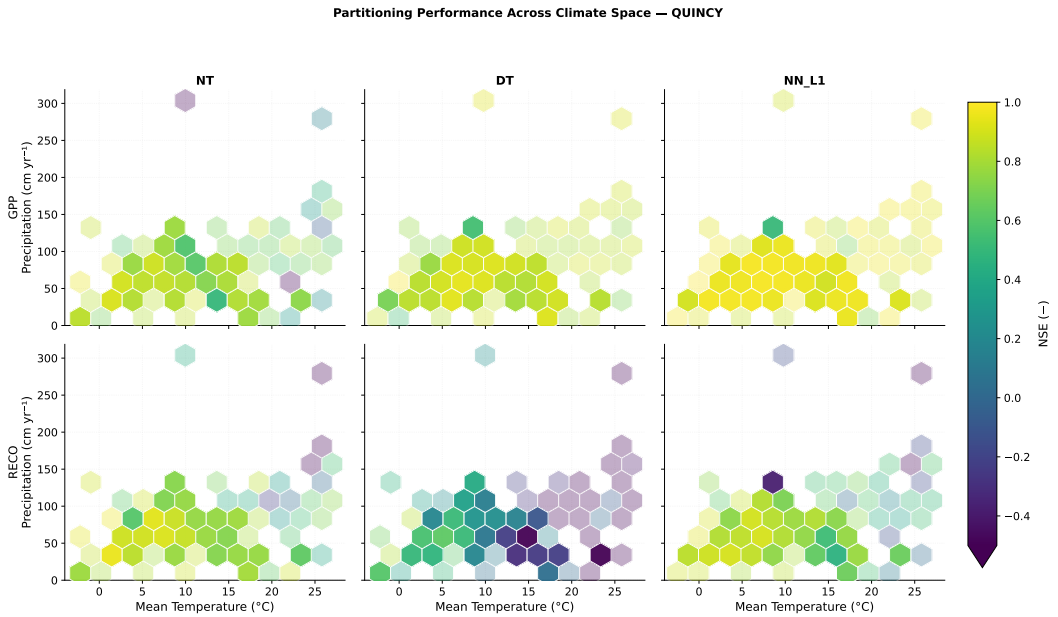


Figure 19: Partitioning performance across climate space (QUINCY). Hexagons show the across-site median of NSE for GPP (top) and RECO (bottom), plotted against site mean annual temperature and precipitation, for NT, DT, and NN_{L1}. NSE for each site is computed once over that site's full multi-year record. Cells aggregating fewer than 3 sites are shown at reduced opacity. Opacity reflects the number of contributing sites, not the number of years of underlying data. The colorbar is clipped at -0.5 and cells with lower median NSE are shown at the minimum color.

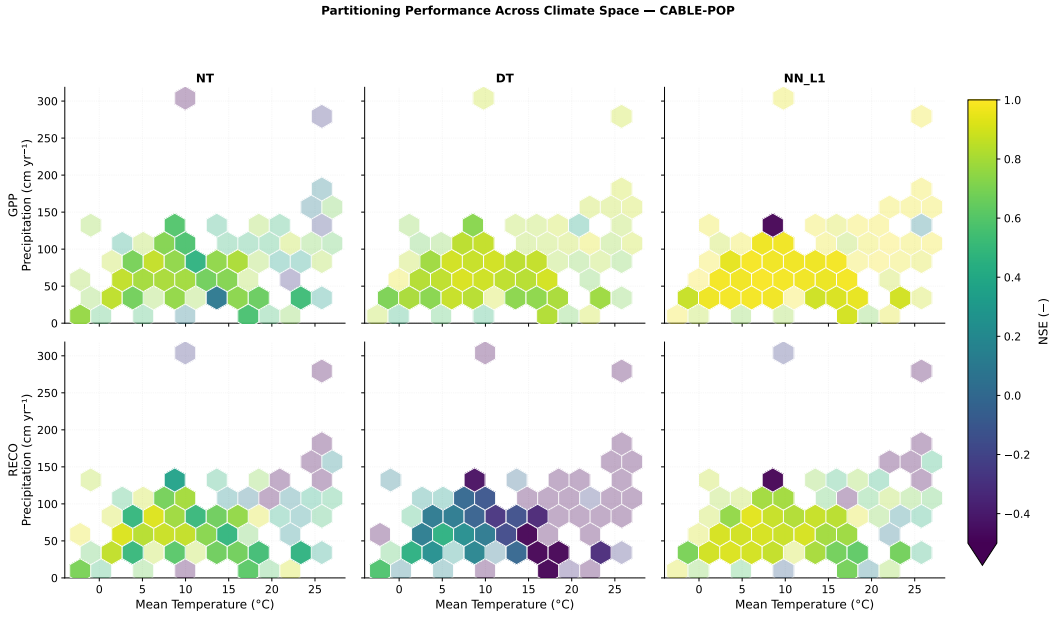


Figure 20: As fig. 19, for Cable-Pop.

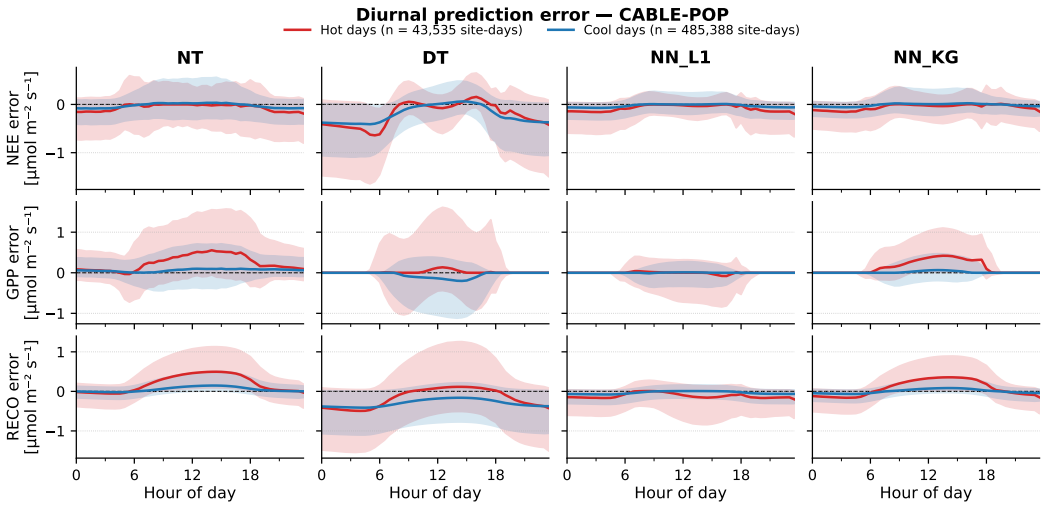


Figure 21: As fig. 6, for Cable-Pop.

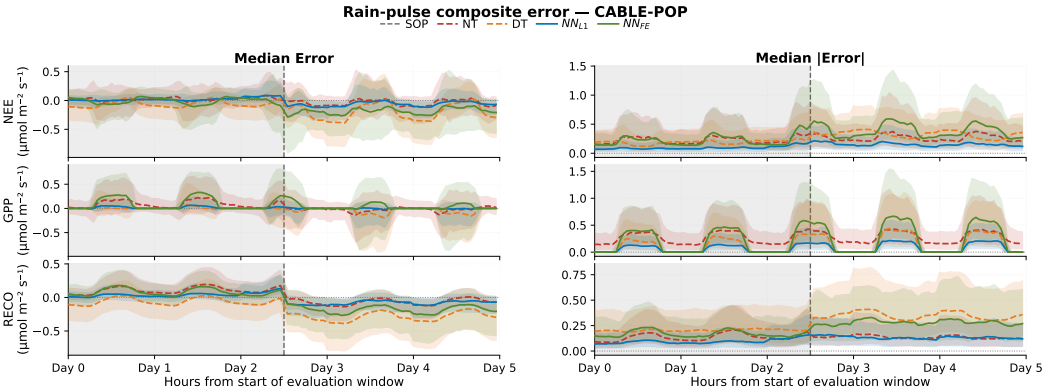
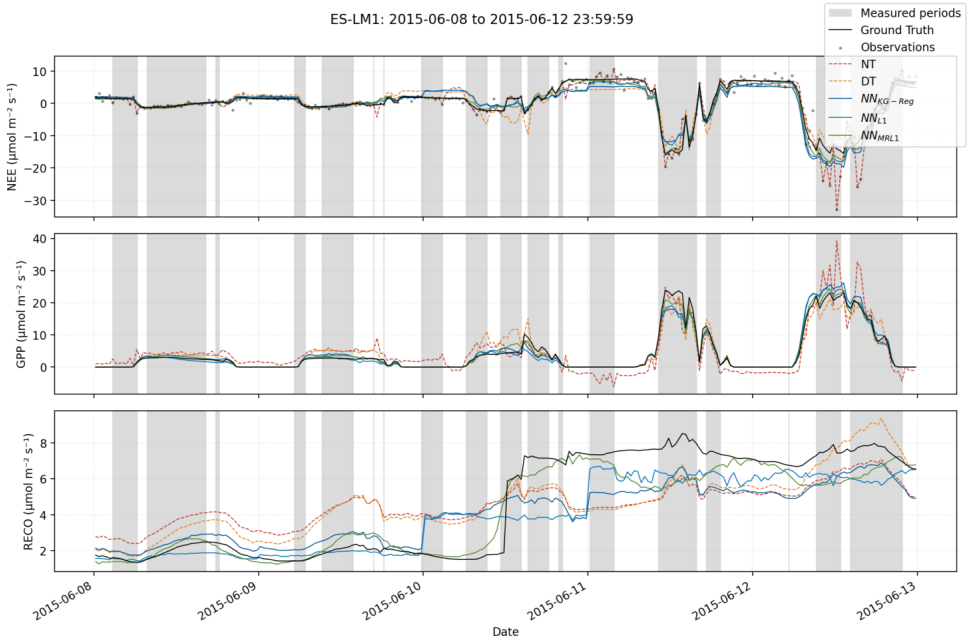
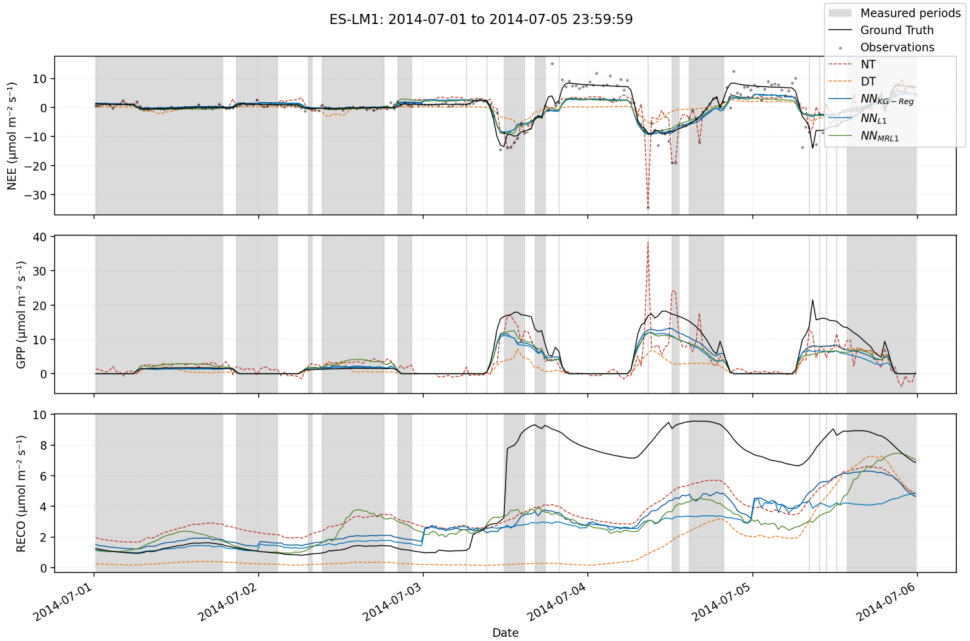


Figure 22: As fig. 7, for Cable-Pop.



(a) ES-LM1, 2015-06-08 to 2015-06-12: $\mathrm{NN_{FE}}$ captures the rise via its moving-window feature.



(b) ES-LM1, 2014-07-01 to 2014-07-05: all methods mostly miss the pulse.

Figure 23: Rain-pulse response at two QUINCY events. Simulated NEE, GPP, RECO with reintroduced noise ("Simulated obs.") and availability masking ("Measured periods," shaded), ground truth, and five methods. (a) and (b) show the mechanism is event-dependent, not a robust fix.



Figure 24: Inter-method spread across climate space. Hexagons show, for each climate cell, the across-site median of the per-timestep standard deviation across the group (NT, DT, DML, NN_{L1}, $\mathrm{NN}_{\mathrm{KG}}$, $\mathrm{NN}_{\mathrm{FE}}$), computed per site and normalized by the standard deviation of NEE in the ground truth (QUINCY, CABLE-POP) or measured record (Measured). As in fig. 19, cells aggregating fewer than 3 sites are shown at reduced opacity, and opacity reflects site count rather than years of underlying data.

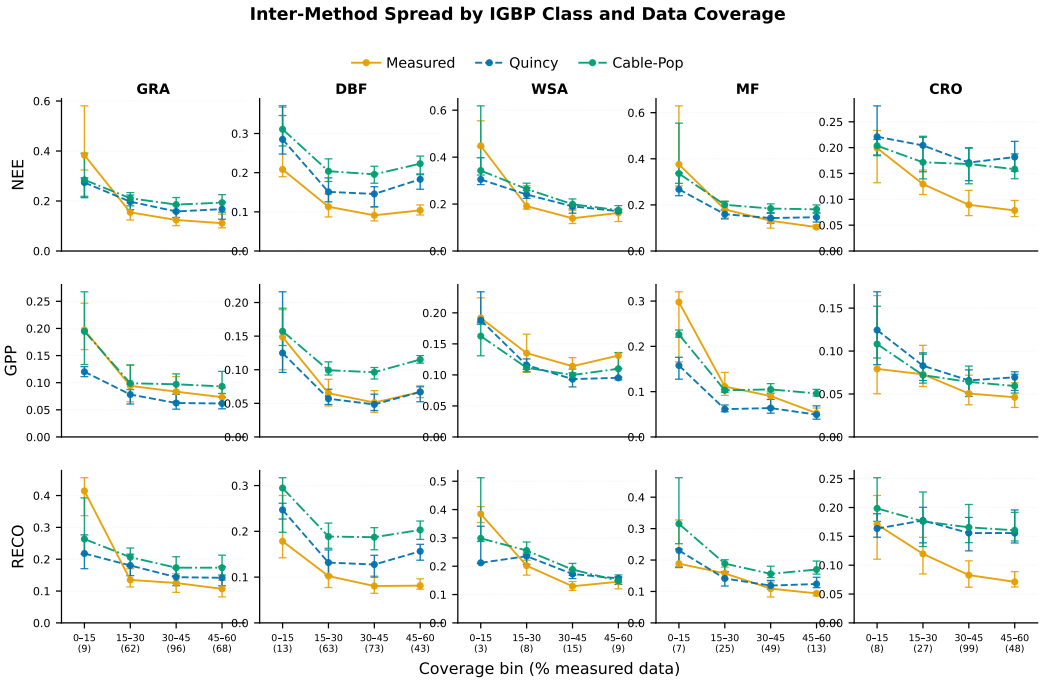


Figure 25: Inter-method spread by IGBP class and data coverage. Normalized spread (median across-site standard deviation across the method group (NT, DT, DML, NN_{L1}, $\mathrm{NN}_{\mathrm{KG}}$, $\mathrm{NN}_{\mathrm{FE}}$), normalized as in fig. 24, but per site-year) for five IGBP classes: grassland (GRA), deciduous broadleaf forest (DBF), woody savanna (WSA), mixed forest (MF), and cropland (CRO). Site counts per bin given in parentheses.

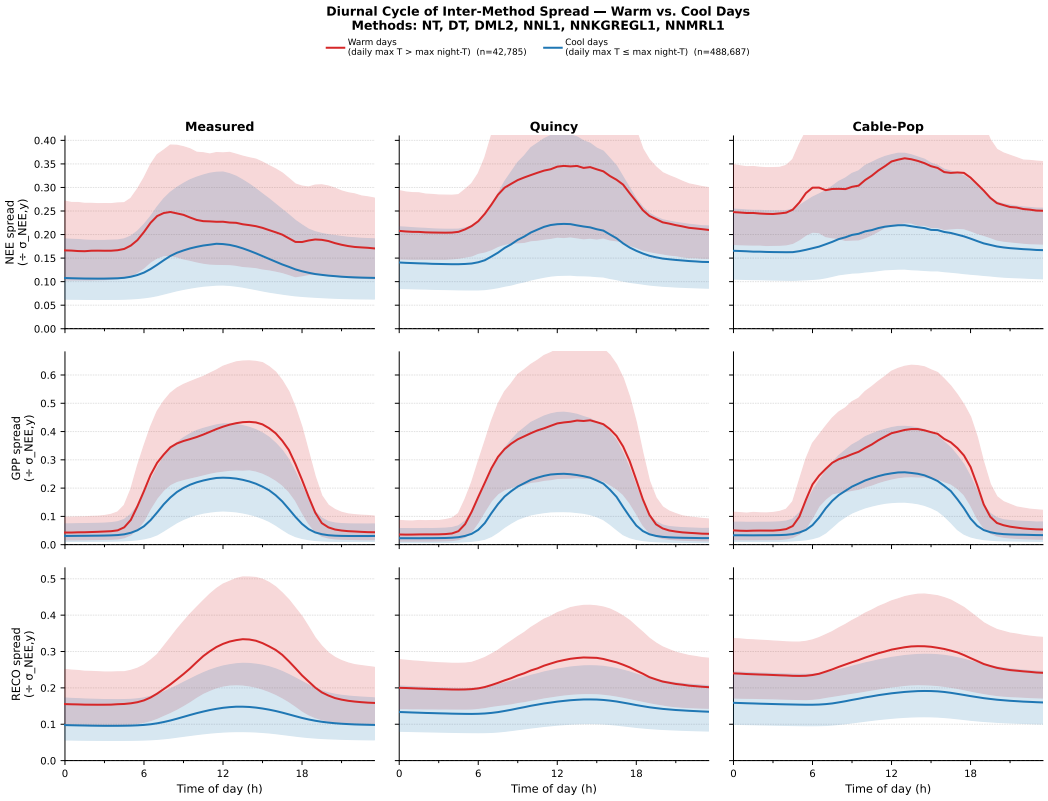


Figure 26: Diurnal cycle of inter-method spread, separated by warm and cool days. Spread is the median of standard deviations across method group (NT, DT, DML, NN_{L1}, $\mathrm{NN}_{\mathrm{KG}}$, $\mathrm{NN}_{\mathrm{FE}}$), normalized as in fig. 24, but per site-year. Days are classified "warm" (daily max temperature exceeds the site-year's max nighttime temperature) or "cool". Shaded bands show IQR across all days.

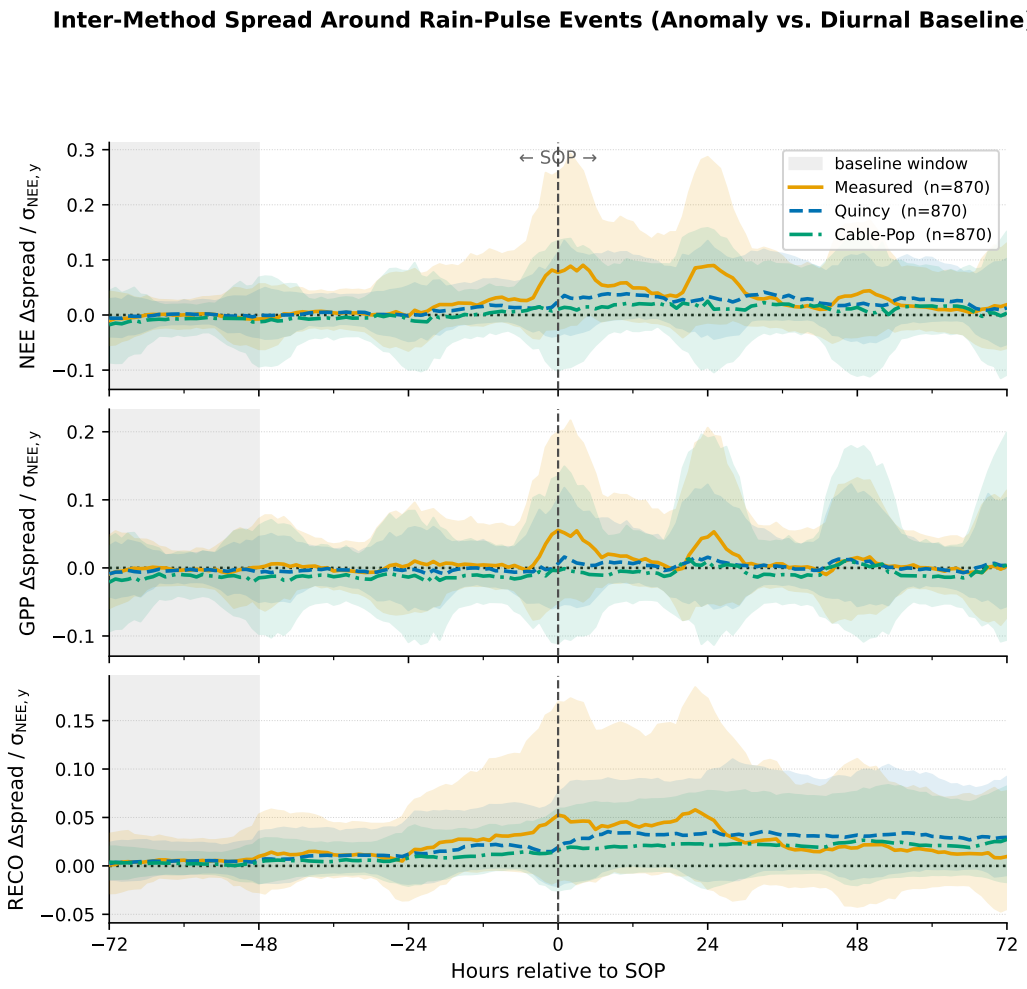


Figure 27: Spread of partitioning methods across rain-pulse events. Standard deviation across method group (NT, DT, DML, NN_{L1}, $\mathrm{NN}_{\mathrm{KG}}$), normalized relative to each event's per-time-of-day diurnal baseline, computed over the 5-day pre-pulse window (SOP -7 to SOP -2 days). Aggregated across n=870 validated rain-pulse events. Lines show the median; bands show the IQR across events. The shaded region marks the portion of the baseline window falling within the plotted ±3-day range. The full 5-day baseline extends earlier than shown.

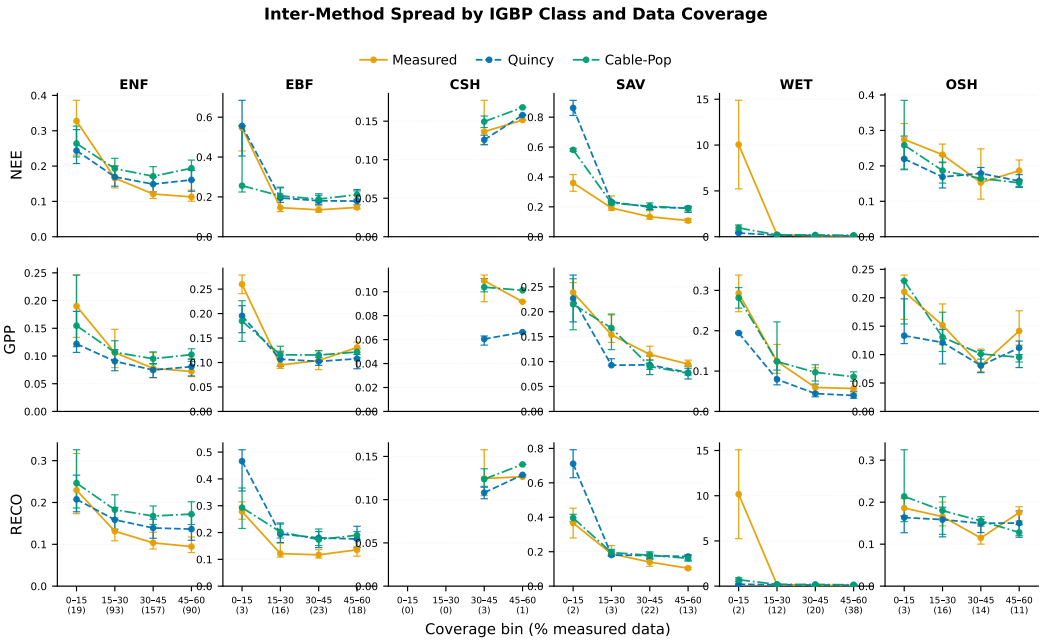


Figure 28: As in fig. 25, but for the second group of IGBP classes: evergreen needleleaf forest (ENF), evergreen broadleaf forest (EBF), closed shrubland (CSH), savanna (SAV), wetland (WET), and open shrubland (OSH).