Ecological Informatics 96 (2026) 103820

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Modeling gross primary productivity across different European ecosystem types: Evaluating the versatility of SARIMAX, XGBoost, and LSTM using ICOS FLUXNET and Sentinel-2 data

Anna Spinosa [a,b,*,1], Karisma Karisma [a,c,1], Marieke A. Eleveld [a,d], Mario Alberto Fuentes-Monjaraz [a], Valeria Mobilia [a], Ulf Mallast [e], Johannes Peterseil [f], Ghada El Serafy [a,b]

[a] Deltares, Delft, The Netherlands
[b] Department of Applied Mathematics, Delft University of Technology, Delft, The Netherlands
[c] Department of Applied Mathematics, University of Twente, Enschede, The Netherlands
[d] Department of Geoscience and Remote Sensing, Delft University of Technology, Delft, The Netherlands
[e] Helmholtz Centre for Environmental Research GmbH - UFZ, Leipzig, Germany
[f] Environmental Agency of Austria, Vienna, Austria

ARTICLE INFO

Dataset link: https://doi.org/10.18160/S6HM-CP8Q, 10.4121/b26f4168-6359-4257-8ef2-3362d6bc6593

Keywords:
Gross primary productivity
Sentinel-2
SARIMAX
XGBoost
LSTM
Incremental learning

ABSTRACT

Predicting Gross Primary Productivity (GPP) is key for understanding ecosystem health and quantifying the global carbon cycle. While data-driven models have shown strong performance in capturing GPP dynamics at specific sites, their ability to generalize across ecosystems without site-specific recalibration remains largely untested. This study addresses this gap by evaluating the applicability of XGBoost and LSTM models in estimating GPP across different European ecosystems. We developed a unified (cross-site) modeling framework that integrates in-situ eddy covariance observations and Sentinel-2–derived vegetation indices using incremental learning. Models' performance was assessed via: (i) site-specific models, developed to capture individual site characteristics, and (ii) cross-site generalization, including evaluation on an independent dataset of unseen ecosystems. SARIMAX is included as a site-specific statistical benchmark for comparison. Our findings indicate that XGBoost consistently outperformed the other models, achieving site-specific R^2 values above 0.90 in forest and grassland ecosystems and an average R^2 of 0.72 across unseen sites (range 0.66–0.78). LSTM exhibited better accuracy in predicting GPP peaks at site-specific level, particularly in cropland and forest ecosystems. At site-level, SARIMAX showed comparable performance to XGBoost but struggled in capturing the rapid temporal variation of GPP. These findings demonstrate the feasibility of a data-driven framework for cross-site GPP monitoring within European flux-tower networks, making a first step toward transferable GPP prediction without site-specific recalibration.

1. Introduction

Gross Primary Productivity (GPP) indicates the amount of carbon dioxide (CO_2) fixed in an ecosystem through photosynthesis (Lu et al., 2024). It represents the largest atmosphere-to-land CO_2 flux and serves as a critical measure in understanding the global carbon cycle (Grace, 2004). To date, the ocean and terrestrial ecosystems, despite differences among biomes, have absorbed about half of the fossil fuel emissions (Integrated Carbon Observation System, 2022), and are thus vital in mitigating and regulating global warming. Climate change and environmental stressors are, however, altering ecosystem functioning. Hence, accurate estimation of GPP is essential for monitoring terrestrial carbon dynamics and ecosystem health.

Traditionally, GPP estimates have relied on in-situ measurements of the total exchange of CO_2 using chamber-based methods or Eddy Covariance (EC) techniques, which provide high-frequency data (Papale et al., 2006). Despite the growing development and increasing number of EC towers within global networks (e.g., FLUXNET, ICOS, eLTER, TERN, etc.), these remain spatially limited and unevenly distributed across the globe. Remotely-sensed vegetation indices (VIs) complement EC point-measurement products by enabling broader spatial monitoring of plant productivity (Robinson et al., 2018; Zhang et al., 2023;

* Corresponding author at: Deltares, Delft, The Netherlands.
E-mail address: anna.spinosa@deltares.nl (A. Spinosa).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ecoinf.2026.103820
Received 1 May 2025; Received in revised form 10 May 2026; Accepted 10 May 2026
Available online 14 May 2026

Astola et al., 2019), vegetation health, growth levels, stress, and other conditions (Zhu et al., 2024). Together, in-situ and remotely sensed data can be used to create long-term and consistent time series data. To quantify GPP consistently across different sites and ecosystem types without site-specific recalibration, a standardized protocol is needed.

Data-driven models, including machine learning (ML) and deep learning (DL), have been increasingly adopted to capture GPP temporal variations and magnitudes, replacing physics-based models, inherently complex and requiring a large number of parameters that are globally available but only at coarse resolution (Lu et al., 2024). Yet, most high spatiotemporal data-driven GPP models are trained and validated at individual sites, and their cross-sites transferability remains largely unexplored. As a result, it remains unclear whether data-driven models can provide GPP estimates across biomes without site-specific recalibration.

Among data-driven algorithms for GPP estimation Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and trees-based models are the most widely used approaches (Liao et al., 2023; Tramontana et al., 2015; Liu et al., 2016). ANNs offer flexible architecture but require careful hyperparameter tuning (Chen et al., 2022). SVMs have shown good performance in GPP estimations (Lee et al., 2020); yet their training is computationally expensive and requires a large amount of memory. Tree-based models are generally computationally less intense and have the advantage of reducing overfitting while effectively handling heterogeneous predictors. Among ML algorithms, tree-based methods, particularly the Extreme Gradient Boosting (XGBoost), have achieved higher accuracy and computational efficiency than ANN, SVM and Random Forest (RF) in modeling GPP and carbon flux (Na et al., 2025; Wang et al., 2023; Liu et al., 2021). In parallel, DL has also been increasingly adopted in ecological and GPP time-series modeling, outperforming SVM, RF and ANN (Lee et al., 2020). Photosynthetic activity exhibits memory effects from prior environmental states, motivating the use of Recurrent Neural Networks (RNNs, a class of DL), designed to be effective in retaining temporal dependencies. Within this class, Long Short-Term Memory (LSTM) networks have shown good performance in modeling GPP, including improved representation of climate-induced productivity extremes (Montero et al., 2024).

Building on these insights, we selected XGBoost and LSTM for our modeling approach. XGBoost was chosen also for its balance of interpretability, accuracy, and computational efficiency, as well as its ability to effectively handle heterogeneous predictors, such as vegetation indices and meteorological variables (Chen and Guestrin, 2016). LSTM was selected because it is particularly suited to capturing complex ecosystem dynamics, reflecting both immediate meteorological conditions and delayed vegetation responses. Additionally, we implemented the Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX), a statistical model successfully applied in ecological applications where seasonality and external environmental drivers play a major role (Zhao et al., 2022). Its interpretability makes it a useful comparative model for assessing the added predictive value of ML and DL methods.

In this study, we evaluate the feasibility of a cross-site incremental learning framework for transferable GPP prediction without site-specific recalibration. To this end, the research was conducted in two phases. In the first phase, we developed site-specific (individual) models for each training site. The three complementary modeling approaches were compared at site-specif level to highlight the added value and limitations of each method. In the second phase, we used the XGBoost and LSTM models to develop cross-site (unified) models. The models were implemented in an incremental learning framework. Rather than retraining the models from scratch, incremental learning allows the models to assimilate data from additional sites (ecosystems) while retaining information learnt from previous sites (ecosystems). This approach is particularly suited given that the ICOS sites are commissioned at different times, and data are accumulated sequentially, reflecting the realistic conditions of an expanding monitoring network. The cross-site models were further evaluated on an independent testing dataset to assess their ability to predict GPP in ecosystems not included in training. This final step was specifically designed to evaluate the spatial transferability of data-driven approaches in the absence of GPP measurements, that is, whether a pre-trained model can be directly applied to a new site using only predictor variables, without the need for site-specific retraining. This work should be considered a proof of concept, aiming to evaluate the feasibility of a unified modeling framework for transferable data-driven ecosystem monitoring.

2. Methods

2.1. Area of interest

Four different ecosystem types were selected for the analysis since we aimed at developing a unified model capable of estimating GPP in various European sites and across ecosystem types. The main criterion for site selection was the availability of a complete time series for at least three years to ensure sufficient data for the model training process. Moreover, sites were chosen at different locations spanning a wide latitude and longitude range to achieve the coverage of a large geographical area.

In total, eight sites were selected (Fig. 1) belonging to four ecosystem types: evergreen needleleaf forest (ENF), deciduous broadleaf forest (DBF), cropland, and grassland. Two sites for each ecosystem type were selected. The eight sites were split into two sets, with one set serving as training sites for building both individual (site-specific) and unified (cross-site) models, and the other set serving as test sites to evaluate the performance of the unified model in predicting GPP on unseen sites. Both sets had the same combination of ecosystem types (Table 1).

2.2. In-situ measurements

The in-situ data were collected from the ICOS portal (Integrated Carbon Observation System, 2025). The ARCHIVE data product for ecosystem measurements at daily resolution was collected, for the period 2017–2023. The ARCHIVE data product contains a set of meteorological variables gap-filled and/or downscaled from the ERA5 dataset (Pastorello et al., 2020). While tower-level observations may be considered more accurate, the ARCHIVE dataset provides consistent and harmonized records across all sites, supporting model reproducibility and transferability. Furthermore, these data have undergone a uniform and standardized quality control process that ensures their reliability (Pastorello et al., 2020). Only the variables that were available for all chosen study sites were used to ensure model reproducibility (Table 2). Those variables include independent meteorological observations (short- and longwave radiation, air temperature, atmospheric pressure, precipitation, wind speed, and vapor pressure deficit), which are broadly available across climate networks, and turbulent flux products derived from direct eddy covariance measurements (sensible heat flux and latent heat flux). The latter typically rely on flux observations and are therefore limited to instrumented sites. These variables were used as predictors to train the models since they represent direct environmental drivers of photosynthesis or its proxies. The target variable for model training was GPP derived from the daytime partitioning method, 50th percentile estimates (GPP_DT_VUT_USTAR50) (Lasslop et al., 2010). This product was selected because the daytime (DT) partitioning method accounts for the effect of vapor pressure deficit (VPD) on the light-response function, thereby reducing biases commonly observed in the nighttime (NT) method, where the VPD effects are typically neglected (Lasslop et al., 2010). Additionally, the "USTAR50" product, representing the median value (50th percentile) of the variable friction velocity (USTAR) threshold (VUT) distribution, was chosen among recommended reference products due to its robustness and

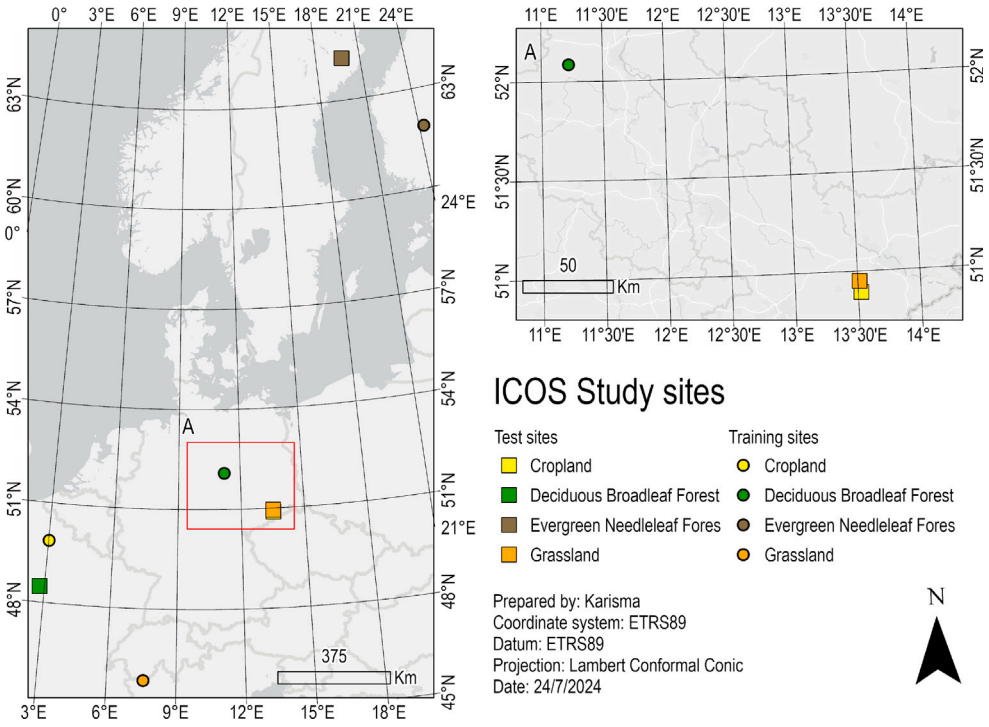

Fig. 1. Map of ICOS stations used in this study. Training sites are marked by squares and test sites by circles. Different ecosystems are indicated by color: cropland (yellow), DBF (green), ENF (brown), and grassland (orange).

Table 1
Training and testing sites.

Group	Country	Name	Ecosystem type	Start date	End date	Records	ICOS data citation
Train	FR	Estrees-Mons	Croplands	26-05-17	15-10-23	2334	Leonard et al. (2025)
	DE	Hohes Holz	DBF	01-01-19	26-09-23	1730	Rebmann et al. (2025)
	FI	Hyytiala	ENF	01-01-18	22-09-23	2370	Mammarella et al. (2023)
	IT	Torgnon	Grasslands	01-01-18	22-09-23	2091	Cremonese et al. (2025)
Test	DE	Klingenberg	Croplands	01-01-18	25-09-23	2093	Bernhofer et al. (2025a)
	FR	Font.Barbeau	DBF	01-01-19	07-09-23	1770	Berveiller et al. (2025)
	SE	Svartberget	ENF	01-01-19	21-10-23	1754	Peichl et al. (2025)
	DE	Grillenburg	Grasslands	24-04-17	28-09-23	2348	Bernhofer et al. (2025b)

Table 2
List of the in-situ variables collected from ICOS.

Variables	Explanation	Unit
H_F_MDS	Sensible heat turbulent flux	$W\,m^{-2}$
LE_F_MDS	Latent heat turbulent flux	$W\,m^{-2}$
LW_IN_F	Incoming (down-welling) longwave radiation	$W\,m^{-2}$
PA_F	Atmospheric pressure	kPa
P_F	Precipitation	$mm\ d^{-1}$
SW_IN_F	Shortwave radiation	$W\,m^{-2}$
TA_F	Air temperature	°C
VPD_F	Vapor pressure saturation deficit	hPa
WS_F	Wind speed	$m\ s^{-1}$
GPP	Gross primary productivity	$gC\,m^{-2}\,d^{-1}$

stability across different temporal aggregation resolutions (Pastorello et al., 2020).

Global variables, time-invariant descriptors, which serve as proxies for the distinct ecosystems of each site, were collected. These variables included elevation, latitude, and longitude. Ecosystem type and season, derived from the months of data collection, were also utilized as categorical explanatory variables. Seasons were categorized as follows: winter (December to February), spring (March to May), summer (June to August), and autumn (September to November) (Chang et al., 2023). Both the type of ecosystem and the season were treated as categorical data, subsequently converted into a binary format using one-hot encoding, a common technique used to represent each category as a separate binary column (1 or 0). This approach facilitates the integration of categorical variables into the analysis and allows their use in machine learning models.

2.3. Remote sensing data

Sentinel-2 Level-2 A data products, resampled at 10 m resolution, were used to derive the VIs, used as a proxy for GPP. Information on the spectral response functions of Sentinel-2 data can be found in Table 3 and on the Sentinel-2 handbook (ESA, 2025).

The vegetation indices derived from Sentinel-2 provide information about leaf pigments, leaf and canopy structure (Frampton et al., 2013). The Normalized Difference Vegetation Index (NDVI) was originally developed to detect vegetation health over a range of latitudes by comparing reflectance in the red and near-infrared bands. Refinements, such as the Modified Normalized Difference Vegetation Index (MNDVI), the Enhanced Vegetation Index (EVI), and the Two-Band Enhanced Vegetation Index (EVI2) (Table 3) were primarily developed for estimating the leaf area index (LAI), defined as the one-sided green leaf area per unit ground surface area, as well as for the retrieval of canopy chlorophyll content. These VIs are derived using bands in the red vegetation maximum chlorophyll absorption (Band 4), the red-edge position (Band 5), and the high reflectance in the near-infrared (NIR) region

Table 3
Spectral responses for each band of Sentinel-2 multispectral instrument.
Source: https://sentinels.copernicus.eu/-/copernicus-sentinel-2c-spectral-response-functions

Band #	Centre λ (nm)	Spectral width $\Delta\lambda$ (nm)	Spatial resolution (m)	Purpose
B1	443	20	60	Atmospheric correction (aerosol scattering).
B2	490	65	10	Sensitive to vegetation senescing, carotenoid, browning and soil background; atmospheric correction (aerosol scattering).
B3	560	35	10	Green peak, sensitive to total chlorophyll in vegetation.
B4	665	30	10	Maximum chlorophyll absorption.
B5	705	15	20	Position of red edge; consolidation of atmospheric corrections/fluorescence baseline.
B6	740	15	20	Position of red edge, atmospheric correction, retrieval of aerosol load.
B7	783	20	20	Leaf Area Index (LAI), edge of the Near-Infrared (NIR) plateau.
B8	842	105	10	LAI.
B8a	865	20	20	NIR plateau, sensitive to total chlorophyll, biomass, LAI and protein; water vapor absorption reference; retrieval of aerosol load and type.
B9	945	20	60	Water vapor absorption, atmospheric correction.
B10	1375	30	60	Detection of thin cirrus for atmospheric correction.
B11	1610	90	20	Sensitive to lignin, starch and forest above ground biomass. Snow/ice/cloud separation.
B12	2190	180	20	Assessment of Mediterranean vegetation conditions. Distinction of clay soils for the monitoring of soil erosion. Distinction between live biomass, dead biomass and soil, e.g. for burn scars mapping.

Table 4
List of vegetation and water indices derived from Sentinel-2 Level-2A products.

Abbreviation	Name	Equation
NDVI	Normalized Difference Vegetation Index	$\frac{NIR-Red_{665}}{NIR+Red_{665}}$
MNDVI	Modified Normalized Difference Vegetation Index	$\frac{NIR_{783}-Red_{705}}{NIR_{783}+Red_{705}}$
EVI	Enhanced Vegetation Index	$\frac{NIR-Red_{665}}{NIR+6Red_{665}-7.5Blue+1}$
EVI2	Two-Band Enhanced Vegetation Index	$\frac{NIR-Red_{665}}{NIR+2.4Red_{665}+1}$
CIr	Red-edge Chlorophyll Index	$\frac{NIR_{783}}{Red_{705}}-1$
LSWI	Land Surface Water Index	$\frac{NIR-SWIR1}{NIR+SWIR1}$
NDII	Normalized Difference Infrared Index	$\frac{NIR-SWIR2}{NIR+SWIR2}$
MNDWI	Modified Normalized Difference Water Index	$\frac{Green-SWIR1}{Green+SWIR1}$

(Band 8), and enhance estimates of leaf chlorophyll concentration and canopy chlorophyll content. Water-sensitive indices – namely, the Land Surface Water Index (LSWI), Normalized Difference Infrared Index (NDII), and Modified Normalized Difference Water Index (MNDWI) – indicate vegetation water content (plant water status), drought and water stress, and open water detection, respectively.

Different VIs may correlate slightly better for GPP for specific ecosystems and regions (Spinosa et al., 2023), and the need for greenness and water-related indices to effectively capture the spatial and temporal pattern of the GPP has been demonstrated (Noumonvi et al., 2019). Greenness-related VIs are indeed more effective during the wet phase of the growing season, while water-content-related indices perform better in the dry season due to the sensitivity of GPP to water availability. Therefore, both types of indices were used in this study. The equations used to derive the different VIs from Sentinel-2 data are presented in Table 4.

Sentinel-2 Level-2 A images were retrieved at the defined locations and time periods. The period corresponded to the time of available in-situ measurements, since those are used as ground truth, while the location corresponded to the coordinates of a reference area assigned around each flux tower of each site. The reference area was used as an approximation of the non-static climatological footprint, the surface area contributing to the measured flux (Pluntke et al., 2023). The buffer area around the flux tower was manually determined through satellite image inspection and defined as a circle area with a certain radius. The coordinates of the flux tower were used as the center of the area. The radius of the buffer area varied for each site to ensure that the retrieved data covered only the desired ecosystem type (Figs. 1 and 2, Table S1 of the supplementary). At first, images within the defined coordinates and period were retrieved. Thereafter, images were filtered by cloud coverage; those with more than 30% cloud coverage were excluded from the analysis. Finally, non-vegetated pixels were removed from the image using the Sentinel-2 Scene Classification Layers (SCL) provided by ESA, which distinguish pixels among twelve classes (e.g., vegetation, snow or ice, defective pixels, etc.) (Copernicus Sentinel Hub, 2025).

After cloud filtering, the VIs values calculated per pixel within the buffer area were averaged to obtain a single daily value. To mitigate the impact of outliers on subsequent analysis, a z-score technique was applied. Data points with z-score equal to or higher than 3 were removed. Linear interpolation was used to estimate missing data points, followed by the application of a Savitzky-Golay filter (Savitzky and Golay, 1964) to smooth the time series and enhance data quality.

2.4. Models

Three different models and their performances were analyzed in this study: SARIMAX, XGBoost, and LSTM.

The AutoRegressive Integrated Moving Average (ARIMA) model, popularized by Box and Jenkins (Box and Jenkins, 1970), has been extensively used to predict trends based on historical data (Fattah et al., 2018; Ariyo et al., 2014). The SARIMAX model extends the ARIMA framework by incorporating both exogenous variables (external predictors, e.g., meteorological drivers) and seasonal components, allowing

for a more accurate representation of periodic environmental influences and climate drivers. SARIMAX has been shown to improve forecasts in the fields of hydrology (Fathi et al., 2019), ecology (Guo et al., 2023), and healthcare (Tolcha, 2023; Kumar et al., 2023). SARIMAX models are usually denoted $SARIMAX(p,d,q)(P,D,Q,s)$ (nomenclature in Table S2, supplementary). The lowercase p,q,d correspond to the number of past observations (lags) used to predict the current value, the number of times the time series is differenced to remove trends and make it stationary (i.e., constant mean and variance over time), and the number of past errors included in the model to predict the current value. The uppercase P,D,Q are the autoregressive, differencing, and moving average terms of the seasonal component s representing the seasonal period (for example, the seasonal period $s = 7$ corresponds to a weekly seasonality in the daily data).

The XGBoost model is an ensemble model that builds decision trees sequentially, with each new tree correcting the residuals of the previous ones through gradient boosting. This approach allows the model to learn complex nonlinear relationships while mitigating overfitting through internal regularization (Chen and Guestrin, 2016). To control the tree's growth and prevent overfitting, which occurs when a model learns patterns specific to the training data, resulting in high accuracy on the training set but poor performance on unseen test data, hyperparameters like the minimum child weight and the maximum tree depth can be tuned, ensuring the tree's splits are meaningful and limiting the model complexity. Additional hyperparameters, including the learning rate *eta*, the regularization term *alpha*, and *gamma*, which sets the minimum loss reduction required to make a split, are tuned to allow for stable and conservative learning. The definition of the XGBoost hyperparameters optimized in this study is provided in Table S2 of the supplementary.

LSTM networks, a deep learning variant of recurrent neural networks, are designed to capture long-term temporal dependencies through gated memory cells that regulate information flow over time (Hochreiter and Schmidhuber, 1997). During training, the loss function is minimized over multiple epochs, the number of times the model iterates over the entire training dataset. Monitoring the training loss across epochs ensures that the model is learning and not overfitting. Adjusting hyperparameters such as the number of LSTM units per layer, which controls the network learning rate, and the number of epochs, can help improve training performance. The definition of the LSTM hyperparameters optimized in this study is provided in Table S2 of the supplementary.

To capture temporal dependencies in the data, lagged values of the predictor variables are included as input features for XGBoost and LSTM. Lag values at the train/test boundary are derived from training observations only, ensuring no future information is used.

Using the SARIMAX, XGBoost and LSTM models, we performed two experiments (Fig. 2). In the first experiment, we evaluated the performances of site-specific (or individual) models. For each site, we applied the three different algorithms and assessed their performances to identify the most effective model. In the second experiment, we constructed a cross-site (unified) model using incremental learning. This approach involved sequentially training the model on data from each site, allowing it to retain knowledge from previous training steps and adapt to new sites. Incremental learning was employed for the XGBoost and LSTM algorithms; SARIMAX was used as a site-specific statistical baseline and was not extended to the unified cross-site framework. The performance of the unified model was then compared with that of the site-specific models. Additionally, the unified model was evaluated on the second set of unseen sites (Table 1) to assess the models' versatility and performance on data they had not encountered during training, thereby allowing us to assess their ability to generalize to new locations without site-specific calibration.

2.4.1. Experimental design 1: individual (site-specific) models

In the first stage of model development, individual sites were analyzed. The performances of the SARIMAX, XGBoost and LSTM models were assessed and compared using predefined evaluation metrics to determine the most effective approach for each site. The employed evaluation metrics were: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). Additionally, the Akaike Information Criterion (AIC) was computed for the SARIMAX model. AIC was only computed for SARIMAX, as this is a likelihood-based parametric model for which the likelihood and number of estimated parameters are explicitly defined. By contrast, AIC does not apply to XGBoost and LSTM, whose training is based on minimizing regularized loss functions rather than maximizing a statistical likelihood.

Before training, the datasets from each site were partitioned chronologically into training and testing subsets with a split ratio of 80:20. The earliest observations were assigned to training and the most recent 20% reserved for testing. This approach ensures consistency in model training and enhances the reliability of the model evaluation process across different ecological contexts. The models were implemented using: (i) the StatsModels library in Python for SARIMAX (SARIMAX, 2025; Seabold and Perktold, 2010), (ii) the XGBoost package in Python for XGBoost (XGBoost, 2025b), and (iii) Keras library in Python for LSTM (LSTM layer, 2025).

Parameter optimization was conducted via a systematic grid-based search to optimize the performance of the individual site models. The parameter values used in the optimization phase, together with a detailed description of the optimized parameters, can be found in Tables S2 and S3, supplementary. For SARIMAX, the parameters were the order (p,q,d), and seasonal order (P,Q,D,s). For XGBoost, the tuned hyperparameters included minimum child weight (mcw), *eta*, *alpha*, *gamma*, maximum depth of a tree (md), and time lags (XGBoost, 2025a). For the LSTM, the adjusted hyperparameters were the LSTM units of each layer, the number of epochs, and the dropout ratio. A shallow network consisting of two LSTM layers, followed by a dropout layer and an output layer, was used. Dropout is a common method to enhance generalization in neural networks by randomly deactivating neurons during training. This prevents dependency on any single neuron and promotes learning of diverse features (Alzubaidi et al., 2021). The training of the LSTM was conducted using the Adam optimizer and mean squared error as the loss function. The batch size for the training process was set at 64.

For the training of site-specific models, all vegetation indices (Table 4) were used along with the in-situ measurement data (Table 2) as explanatory variables. The response variable was GPP. Global variables (location of the site, season, etc.) were excluded from the training dataset for the individual model. The rationale for this exclusion was that the values for global data remained constant within the same site, not providing any additional insight to enhance model performance. Additionally, the study explored the feature importance of the XGBoost model to identify which indicators most significantly influenced the GPP prediction.

2.4.2. Experimental design 2: unified (cross-site) models

After constructing individual models, the next phase involved building a unified model to explore the capabilities and adaptability of the unified model to generalize GPP prediction across different ecosystem types. The unified model was developed using an incremental learning framework, which entailed continuously training the model with new data. With this approach, adding new sites to the model did not require starting training from scratch; instead, the model utilized the knowledge previously acquired from earlier training sessions. This strategy also addressed the challenge posed by the varying time periods of data availability from different sites.

Data for each site were split chronologically into 80:20 ratios for training and testing, respectively. The model was initially trained using

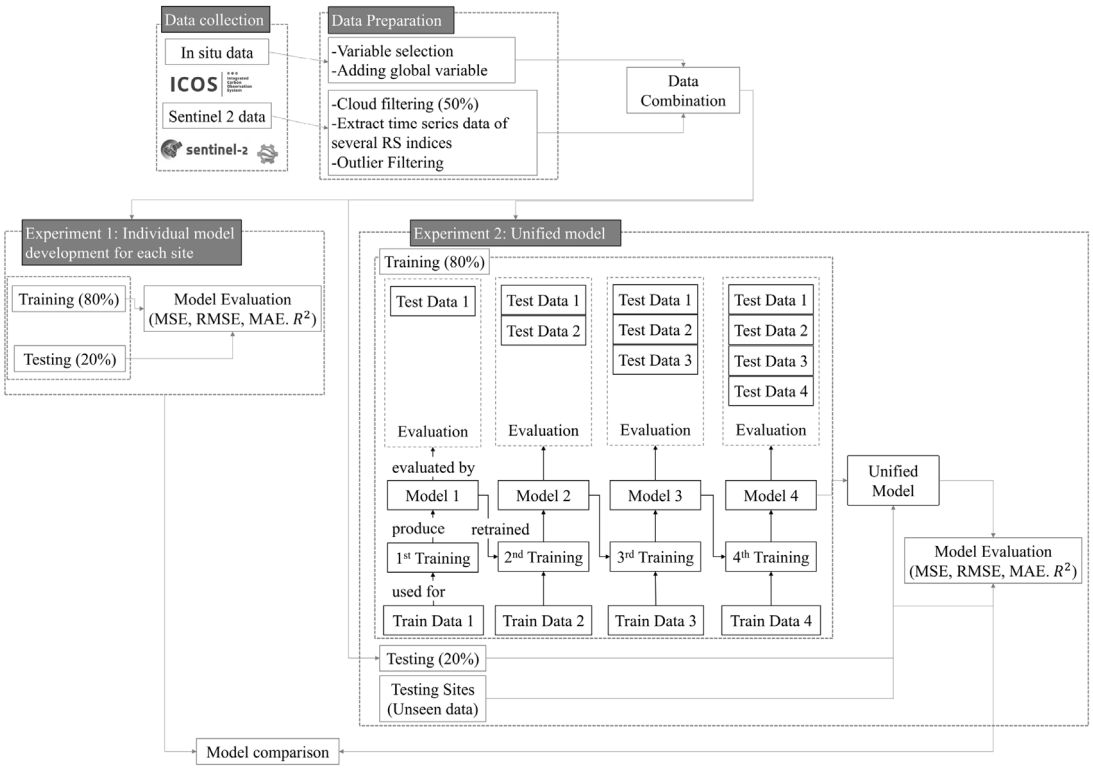

Fig. 2. Flowchart of the methodology, including the training process of the unified model used for XGBoost and LSTM.

data from the first site. Subsequently, the pre-trained model underwent further training with data from the second site, and so on. The training process of the unified model is illustrated in Fig. 2.

Initially, training data from the first site (Train Data 1) were used to produce the first model (Model 1). This model was then evaluated using test data from the same site (Test Data 1). Following the evaluation, Model 1 was retrained using training data from the second site (Train Data 2), resulting in Model 2. Model 2 was evaluated using both the test data from the second site (Test Data 2) and the test data from the first site (Test Data 1) to check for consistency and improvement. This process continued through the 4th training session, culminating in the development of Model 4. Model 4, serving as the unified model, was then assessed to determine the performance achieved through this retraining approach.

The hyperparameters for each algorithm were initialized during the first training, and these settings remained constant throughout the training process. The selection of the hyperparameters was based on the optimal parameters identified through the tuning process of the individual models. The selection was primarily guided by a majority vote from the best parameters across all sites; although this was not the exclusive method used, there was also an exploration of parameters to enhance performance. During the exploratory phase, global variables were utilized as features to determine whether their inclusion could enhance the model's performance. The unified model results presented are those trained without global variables for XGBoost and LSTM, since their inclusion did not enhance performance (Table S6 and Table S8, supplementary).

Incremental learning was employed for constructing unified models using XGBoost and LSTM. The SARIMAX model was not extended to a unified cross-site configuration because it is designed to be fitted to a single continuous time series. As a result, it does not support incremental learning, since incorporating new data typically requires retraining the model from scratch due to the tight coupling of the model's state with the entire time series. Box and Jenkins (1970). SARIMAX was therefore used only as a site-specific statistical baseline.

3. Results

3.1. Individual models

Table 5 presents the accuracy of the estimated GPP. XGBoost exhibits the best performance in terms of evaluation metrics compared to the SARIMAX and LSTM models, showing the highest R^2 and lowest MSE, RMSE, and MAE. For the sites of Hohes Holz (DBF), Hyytiala (ENF) and Torgnon (grassland) an R^2 of 0.91 was achieved. The lowest GPP prediction accuracy (R^2 =0.73) was observed at Estrees-Mons (cropland), due to the sudden interannual variability (e.g., GPP values dropped drastically from around 21.5 $gC\,m^{-2}\,d^{-1}$ on 6 July 2023 to around 1.1 $gC\,m^{-2}\,d^{-1}$ on 14 July, before increasing again on 14 August to 15.9 $gC\,m^{-2}\,d^{-1}$ (Fig. 3)). Among all sites, lowest MSE, RMSE, and MAE values were observed at Torgnon. Model-specific parameter settings are reported in Table S4, supplementary, for each model and site.

3.2. Unified models on training sites

3.2.1. XGBoost performances

Table 6 reports the performance metrics of XGBoost during the training phase of the unified model, using the hyperparameters detailed in Table S5, supplementary. A pattern emerged in the evaluation of models across multiple training sessions. After each model was retrained with data from a subsequent site, performance on previously trained sites decreased. For example, at Estrees-Mons, the MSE increased from 5.73 $(gC\,m^{-2}\,d^{-1})^2$ after the first training session to 9.83 $(gC\,m^{-2}\,d^{-1})^2$ after the second training session, before settling down to 6.35 $(gC\,m^{-2}\,d^{-1})^2$ in the final session. The R^2 followed a similar trend,

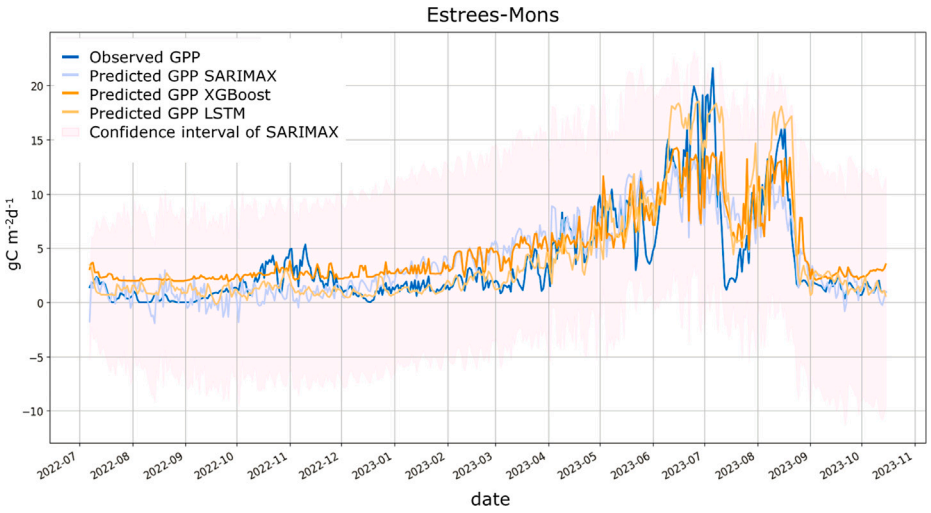

Fig. 3. Predicted vs. Observed Value of GPP at Estrees-Mons. Observed GPP is shown in dark blue. Predicted values from SARIMAX, XGBoost, and LSTM, light blue, orange and yellow, respectively. The pink shaded region denotes the confidence interval of the SARIMAX model.

Table 5
Performance metrics by site and method. RMSE and MAE are expressed in $gC\,m^{-2}\,d^{-1}$; MSE in $(gC\,m^{-2}\,d^{-1})^2$. AIC is reported only for SARIMAX models.

Site	MSE	RMSE	MAE	R^2	AIC	Method
Estrees-Mons	6.59	2.57	1.91	0.65	6527.67	SARIMAX
	5.11	**2.26**	1.84	**0.73**	–	**XGBoost**
	6.07	2.46	**1.66**	0.68	–	LSTM
Hohes Holz	3.65	1.91	1.40	0.88	4708.79	SARIMAX
	2.73	**1.65**	**1.10**	**0.91**	–	**XGBoost**
	4.03	2.01	1.30	0.86	–	LSTM
Hyytiala	1.09	1.04	0.79	0.90	3304.35	SARIMAX
	1.00	**1.00**	**0.63**	**0.91**	–	**XGBoost**
	1.69	1.30	0.89	0.85	–	LSTM
Torgnon	0.95	0.98	0.71	0.89	4083.63	SARIMAX
	0.79	**0.89**	**0.52**	**0.91**	–	**XGBoost**
	0.97	0.99	0.59	0.89	–	LSTM

Table 6
XGBoost incremental training performance metrics by site and training round. RMSE and MAE in $gC\,m^{-2}\,d^{-1}$; MSE in $(gC\,m^{-2}\,d^{-1})^2$.

Site	Metrics	1st Train	2nd Train	3rd Train	4th Train
Estrees Mons	MSE	5.73	9.83	7.28	6.35
	RMSE	2.39	3.14	2.70	2.52
	MAE	1.91	2.31	1.95	1.73
	R^2	0.69	0.47	0.61	0.66
Hohes Holz	MSE	–	4.29	3.77	4.07
	RMSE	–	2.07	1.94	2.02
	MAE	–	1.57	1.29	1.33
	R^2	–	0.86	0.87	0.86
Hyytiala	MSE	–	–	2.76	2.31
	RMSE	–	–	1.66	1.52
	MAE	–	–	1.20	1.03
	R^2	–	–	0.75	0.79
Torgnon	MSE	–	–	–	3.05
	RMSE	–	–	–	1.75
	MAE	–	–	–	1.06
	R^2	–	–	–	0.65
	Average MSE				3.95
	Average RMSE				1.95
	Average MAE				1.29
	Average R^2				0.74

indicating a loss of predictive accuracy, though slightly improved in subsequent sessions. By the final training session, despite some recovery in metric scores at specific sites like Estrees-Mons and Hyytiala, where the MSE improved to 6.35 $(gC\,m^{-2}\,d^{-1})^2$ and 2.31 $(gC\,m^{-2}\,d^{-1})^2$, respectively, the overall model performance declined in comparison to individual models. The average MSE across sites by the last training was 3.95 $(gC\,m^{-2}\,d^{-1})^2$, the average RMSE was 1.95 $gC\,m^{-2}d^{-1}$, the average MAE was 1.29 $gC\,m^{-2}d^{-1}$, and the average R^2 was 0.74.

Fig. 4 displays the feature importance of each individual model, showing the five most important features based on the relative value of the gain metrics. The latent heat turbulent flux contributes to the highest gain value at Hohes Holz and Estrees-Mons, air temperature (TA) emerged as the most significant feature at Hyytiala, whereas at Torgnon, CIr demonstrated the highest gain value.

Table 7
STM incremental training performance metrics by site and training round. RMSE and MAE in $gC\,m^{-2}\,d^{-1}$; MSE in $(gC\,m^{-2}\,d^{-1})^2$.

Site	Metrics	1st Train	2nd Train	3rd Train	4th Train
Hohes Holz	MSE	7.57	9.18	21.22	6.90
	RMSE	2.75	3.03	4.61	2.63
	MAE	1.87	2.02	3.01	1.69
	R^2	0.74	0.69	0.28	0.77
Estrees Mons	MSE	–	9.77	15.86	7.41
	RMSE	–	3.13	3.98	2.72
	MAE	–	2.01	2.70	2.02
	R^2	–	0.48	0.15	0.60
Torgnon	MSE	–	–	1.10	14.89
	RMSE	–	–	1.05	3.86
	MAE	–	–	0.61	3.34
	R^2	–	–	0.87	−0.70
Hyytiala	MSE	–	–	–	2.35
	RMSE	–	–	–	1.53
	MAE	–	–	–	0.99
	R^2	–	–	–	0.79
	Average MSE				7.89
	Average RMSE				2.69
	Average MAE				2.01
	Average R^2				0.36

3.2.2. LSTM performances

Table 7 summarizes the LSTM's performance metrics during the training phase of the unified model, obtained using the hyperparameters detailed in Table S5, supplementary. The site training sequence differed from XGBoost (alphabetical order), since modifying the training sequence improved the LSTM's performance (for a direct comparison using the alphabetical order, see Table S7, supplementary).

The site order influence was further evidenced by the performance degradation observed when testing data from previously trained sites. Specifically, the model showed worsening metrics in subsequent training phases at Hohes Holz and Estrees-Mons. For example, at Hohes Holz the model experienced an increase in MSE, RMSE, and MAE, with the third training yielding an MSE of 21.22 $(gC\,m^{-2}\,d^{-1})^2$ and a low R^2 of 0.28. Similarly, at Hyytiala it exhibited poor metrics during the third training phase. During the fourth training phase, Torgnon's metrics contrasted sharply to the other sites, with MSE escalating from 1.10 $(gC\,m^{-2}\,d^{-1})^2$ to 14.89 $(gC\,m^{-2}\,d^{-1})^2$ and the R^2 turning negative, indicating poor performance, as discussed in Chicco et al. (2021).

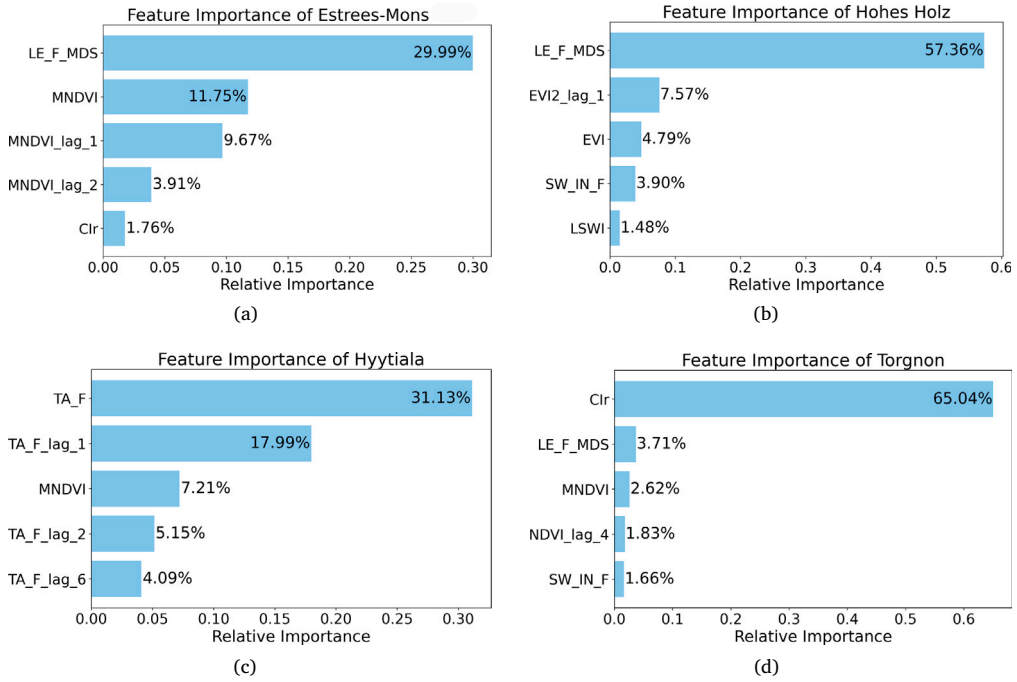

Fig. 4. Feature Importance Scores for XGBoost Models Across Different Sites. For details on specific variables refer to Tables 2 and 4 and Table S2, supplementary.

3.3. Unified models on testing sites

Table 8 compares the performance of the unified models on the unseen testing sites. The XGBoost model achieved higher performance than the LSTM model. The XGBoost average performance metrics across these unseen sites reinforce the model's moderate adaptability, with an MSE of 4.55 $(gC\,m^{-2}\,d^{-1})^2$, RMSE of 2.11 $gC\,m^{-2}\,d^{-1}$, MAE of 1.42 $gC\,m^{-2}\,d^{-1}$, and an R^2 value of 0.72.

3.3.1. XGBoost performances

XGBoost model achieved an R^2 value of 0.78 and 0.77 at the sites of Fontainebleau-Barbeau (DBF) and Svartberget (ENF), respectively, indicating good performance and better generalization (Table 8). This aligns with the earlier trend depicted in Table 6, where the model improved after adapting to similar ecosystems during training, achieving an R^2 value of 0.86 and 0.79 in the fourth training phase for the sites of Hohes Holz and Hyytiala, respectively. The model, however, showed lower performance at the Grillenburg (grassland) site ($R^2 = 0.66$) and at the Klingenberg (cropland) site ($R^2 = 0.68$). Still, the model showed its moderate adaptability at the Klingenberg site (cropland), showing an improvement of performance when compared with Estrees-Mons (cropland), at which the model initially struggled during the training phase.

3.3.2. LSTM performances

Despite poor average evaluation metrics during the training phase (Table 7), the LSTM model showed improved performance during the testing phase (Table 8), achieving an average R^2 of 0.66 across unseen sites. Similar to XGBoost, the LSTM performed better for deciduous broadleaf and evergreen needleleaf forest ecosystem types with an R^2 of 0.76 at Fontainebleau-Barbeau and 0.73 at Svartberget.

Figs. 5–6 show observed and predicted GPP at Klingenberg and Svarberget. The corresponding plots for the remaining sites are provided in the supplementary (Figures S3 and S4). At Klingenberg, XGBoost performed best with an R^2 value of 0.68 and lowest errors. However, it struggled to predict the highest GPP peaks. LSTM overestimated GPP in September. Both XGBoost and LSTM produced occasional negative predicted values, highlighting the need for non-negativity constraints in future model configurations.

Table 8
Unified model performance metrics for the unseen test sites. RMSE and MAE in $gC\,m^{-2}\,d^{-1}$; MSE in $(gC\,m^{-2}\,d^{-1})^2$.

Site	MSE	RMSE	MAE	R^2	Methods
Klingenberg	**5.67**	**2.38**	**1.43**	**0.68**	**XGBoost**
	7.34	2.71	1.80	0.59	LSTM
Fontainebleau-Barbeau	**4.63**	**2.15**	**1.51**	**0.78**	**XGBoost**
	4.99	2.23	1.60	0.76	LSTM
Svartberget	**2.74**	**1.66**	1.15	**0.77**	**XGBoost**
	3.16	1.78	**1.10**	0.73	LSTM
Grillenburg	**5.15**	**2.27**	**1.60**	**0.66**	**XGBoost**
	6.59	2.57	1.78	0.56	LSTM
Average XGBoost	**4.55**	**2.11**	**1.42**	**0.72**	**XGBoost**
Average LSTM	5.52	2.32	1.57	0.66	LSTM

At Svartberget, XGBoost and LSTM showed strong performances, with R^2 values of 0.77 and 0.73, respectively. XGBoost overestimated low GPP values from October 2021 until April 2024.

4. Discussion

In this study, we assessed and compared the performances of three models for time series analysis: the SARIMAX model, a statistical method, the XGBoost model, a machine learning approach, and the LSTM, a deep learning model, for predicting GPP at different sites belonging to different ecosystem types.

4.1. Individual models

Among the three methods, XGBoost consistently delivered the best performance metrics across all sites. SARIMAX outperformed LSTM in predicting GPP at Hohes Holz, Hyytiala, and Torgnon (Table 5), demonstrating that DL methods are not always superior to classical statistical approaches. Although at those sites, the performance of SARIMAX was closely matched with that of XGBoost, the model struggled with rapid temporal changes and tended to underestimate GPP. Consistency between the AIC and the distribution-free error metrics (RMSE, MAE, and

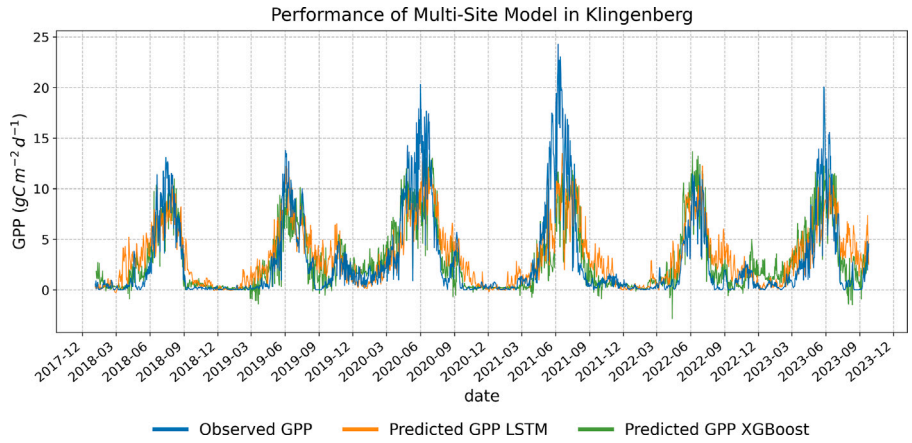

Fig. 5. Predicted vs. observed values of GPP from Klingenberg. Observed GPP in blue, XGBoost predictions in green, and LSTM predictions in orange.

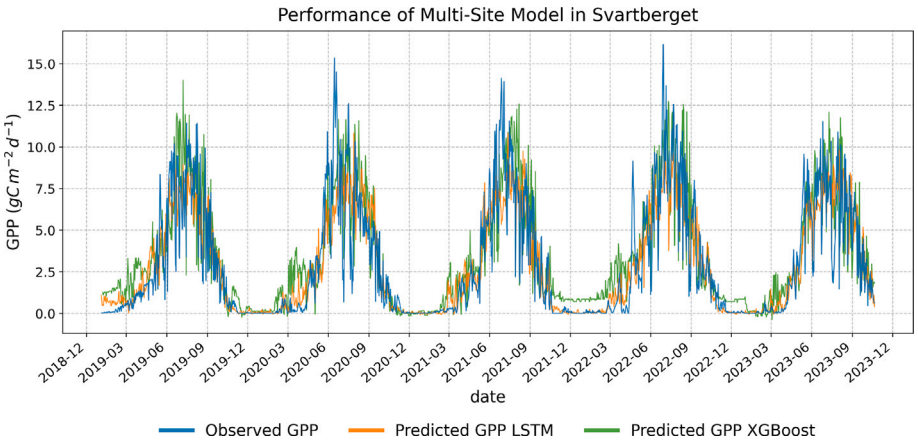

Fig. 6. Predicted vs. observed values of GPP from Svartberget. Observed GPP in blue, XGBoost predictions in green, and LSTM predictions in orange.

R^2) was observed across sites. A significant drawback of SARIMAX was its longer computational time. For instance, the SARIMAX estimation process for the Hyytiala site took about 31 min, whereas XGBoost completed the same task in approximately 4 s (evaluated on an Intel UHD Graphics 620 with 3.9 GB shared GPU memory). This advantage of XGBoost results from its gradient-boosting strategy, which enables faster convergence by incrementally correcting errors with each tree addition, rather than optimizing all parameters simultaneously (Chen and Guestrin, 2016). XGBoost showed better overall performance; however, at some sites, it tended to overestimate GPP during the low-productivity period and underestimate GPP during the high-productivity period. LSTM performed competitively. Its GPP-predicted values tended to be smoother than those from XGBoost and SARIMAX, suggesting LSTM's better generalization capability, particularly in capturing seasonality. While smoothness might be advantageous for constructing a global model that generalizes well across different sites, it is crucial to balance it with the need for accurate GPP predictions. Additionally, LSTM was the only model able to capture strong productivity peaks.

Models tended to underestimate GPP during low-productivity periods, occasionally producing negative values of GPP. This is a clear model limitation. To ensure that the outputs remain ecologically realistic, future implementation should incorporate a non-negativity constraint, for example through output transformations or a positive-only objective function. The latter was not implemented here to preserve full model comparability and evaluate the intrinsic behavior of the unconstrained modeling framework.

Across sites, the predictive performance of the individual models at the Hyytiala evergreen needleleaf forest site showed comparable results to those achieved by Cai et al. (2021) (R^2 of 0.89) and by Wang et al. (2021) (R^2 of 0.76), which utilized a Light Use Efficiency (LUE) modeling approach. Models showed similar results in deciduous broadleaf forest. Lower error metrics were observed at Torgnon, however, these results may not fully reflect the relative effectiveness of the models due to the larger range of GPP values of other training sites. High values of GPP, especially in croplands, were expected and align with observed carbon fixation rates in productive agricultural settings, as documented by Hu et al. (2024). The larger the value to predict, the greater the potential squared error (Tiwari, 2022). This phenomenon accounts for the higher MSE observed at Estrees-Mons compared to other sites.

Estrees-Mons, an "arable crops" site characterized by rotational management practices, posed the greatest challenge for all three models. Rotational management practice introduces interannual variability in vegetation growth patterns and, consequently, in GPP dynamics. Each crop type has distinct growing and harvesting periods, resulting in variable timing and magnitude of productivity peaks across years. For instance, the unusual GPP pattern observed in 2023 likely reflects the harvesting time of the crop, followed by the subsequent planting of fast-growing crops, explaining the quick recovery in GPP in August. A similar pattern has been observed in previous years but remains unexplained by the in-situ explanatory variables (e.g., air temperature). SARIMAX model failed in properly capturing this drop, and while XGBoost showed better overall performance, it tended to overestimate GPP during the low-productivity period and underestimate GPP during the high-productivity period. Among the tested methods, LSTM captured the GPP pattern more accurately, predicting both the significant lows and the subsequent high peaks. This was expected, since the ARIMA model requires stationarity, achieved by removing the trends using (seasonal) differencing. This makes the model more prone to

miss peaks (or highly non-linear patterns). In contrast, NN models are capable of learning non-linear relationships and have a 'longer memory' than SARIMAX, which enables them to better identify the peaks. Slightly better performance of the LSTM in predicting GPP extremes was also shown by Montero et al. (2024). Researchers (Xing et al., 2025; Du et al., 2022; Yuan et al., 2015) have demonstrated that accurate cropland-GPP estimation requires accounting for crop type-specific characteristics, such as photosynthetic capacity, growth patterns, and harvesting time. Incorporating such crop-type information could explain croplands' high GPP range and reduced natural variability, and is therefore suggested to enhance the model performance.

4.2. Unified models

After constructing and evaluating individual models, we built unified (cross-site) models to test their adaptability to generalize across ecosystems by sequentially adding data from multiple sites via incremental learning. Both the XGBoost and LSTM unified models showed lower performance than those on individual sites, indicating the challenge of capturing ecosystem-specific GPP dynamics within a single cross-site model.

Under the unified incremental framework, XGBoost outperformed LSTM on both the training and unseen test sites. Yet, the performances of the unified model were lower than those of the individual one. During incremental learning training, the performance on the earliest-trained site (i.e. Estrees Mons) declined as new sites were added (Table 6), suggesting that the model adapted to new site data while losing some of its predictive capabilities. This phenomenon could be indicative of overfitting to the new site data or of an inability of the model to generalize effectively across different environmental conditions (Wang et al., 2023). Successive training sessions, though aimed at enhancing the model's robustness, might have inadvertently introduced complexities that reduced the model's performance on previously well-modeled sites. In future studies, different methods like the Light Gradient-Boosting Machine (LightGBM) with stronger regularization can be used to improve the model results (Cui et al., 2021; Ke et al., 2017). Additionally, the model produced negative GPP predictions under near-zero GPP conditions, such as during winter. This behavior likely results from statistical noise, since the model generates unconstrained continuous outputs that may overshoot below zero. In future developments, predictions should be constrained through transformations or positive-only objective functions to better reflect the biophysical nature of GPP. When tested on unseen sites, the XGBoost unified model demonstrated moderate generalization, with an average R^2 of 0.72. Variability in performance metrics, however, highlights the need to further investigate and fine-tune the model to enhance its robustness and consistency across different ecological settings.

LSTM showed sensitivity to the order in which sites were introduced during incremental learning. Performance fluctuated across training phases, with models trained on early sites (i.e., Hohes Holz) experiencing decline in performances when new sites were added. Model instability indicated its susceptibility to catastrophic forgetting, where it fails to retain previously learned information when exposed to new data (Jung et al., 2016). Despite recovery in performance during the fourth training phase, the need for strategies such as parameter freezing (Ede et al., 2022), which has been shown to preserve learned information and prevent overfitting to new data, or regularization, which might mitigate catastrophic forgetting in neural networks (Khatib and Karray, 2019), was highlighted and therefore further investigation is suggested. When tested on unseen sites, the model reached an average R^2 of 0.66.

Across ecosystems, the unified models showed consistent patterns. XGBoost and LSTM models performed best at deciduous broadleaf forest and evergreen needleleaf forest sites ($R^2 > 0.76$), where the seasonal cycle of GPP could be more effectively captured (Lai et al., 2025). Both LSTM and XGBoost require improvements for better predictions in grassland sites. Although the unified models well captured seasonal GPP trends, they also showed limitations. Those are not solely model-driven but also reflect the ecosystem-specific processes not fully represented by the current predictors, which do not include direct drivers of GPP as snow cover duration, soil properties, or species-specific physiological responses. For instance, while the sites of Torgnon and Grillenburg (grasslands) are characterized by similar climate conditions, their different elevation (2160 m Torgnon, 385 m Grillenburg) leads to different snow cover periods, which affects the length of the carbon uptake period (Galvagno et al., 2013; Rossini et al., 2014). Likewise, carbon uptake in cropland and grassland is highly dependent on the nutrient availability and growing season length (Galvagno et al., 2013; Owen et al., 2007), while species-specific responses to late spring frost affect forests like Fontainebleau-Barbeau (oak forest) and Hohes Holz (beech forest), differently, with oak being more resilient than beech forest (Rubio-Cuadrado et al., 2021). Owen et al. (2007) further demonstrated that annual GPP in deciduous forest sites generally decreases from south to north, primarily due to shorter growing seasons, a pattern evident in the Fontainebleau-Barbeau and Hohes Holz sites. The fact that the addition of geographic coordinates did not improve model results suggests that air temperature and precipitation already covered the relevant gradient in the growing environment.

To expand model applicability and robustness, the inclusion of soil information, such as soil moisture and evapotranspiration, CO_2 trends, fertilization rates, and drought indicators, among others, should be incorporated as they may provide additional information on the water status and plant physiology that can affect photosynthetic activities. Satellite-based products such as the solar-induced fluorescence (SIF) from Tropospheric Monitoring Instrument (TROPOMI) (Chen et al., 2025) or soil moisture from the Soil Moisture Active Passive Mission (SMAP) (Purdy et al., 2018) and ERA5 dataset can provide more direct and physiologically relevant information on vegetation functioning and plant water stress, not fully captured by precipitation or water and greenness indices alone. Further work may explore the integration of such additional and complementary satellite-based products, whilst accounting for the different temporal and spatial resolutions among the datasets and ensuring the reliability of upscaled estimates and the development of harmonized products. Downscaled products, such as TroDSIF, achieve a resolution of 500 m for 16-day composites (Chen et al., 2025). Higher-resolution SIF observations from NASA's Orbiting Carbon Observatory (OCO)-2 and OCO-3 missions (~ 1–2 km) are available globally but with spatial discontinuities due to narrow swaths. The upcoming ESA FLEX mission (launch planned for 2026) will further enhance SIF monitoring by providing spatially continuous observations at 300 m resolution. Similarly, soil moisture products from the Copernicus Land Monitoring Service (1 km) and the SMAP/Sentinel-1 synergy (1–3 km) (Das et al., 2018) offer detailed spatial information relevant for ecosystem-scale GPP estimation. Yet, the integration of these heterogeneous datasets with in-situ or Sentinel-2 measurements will require careful temporal and spatial harmonization to ensure the consistency and reliability of upscaled GPP estimates.

Despite currently relying only on environmental variables and proxies of leaf water content and greenness, the models still demonstrated promising potential in predicting GPP values across different ecosystem types, using a limited and relatively simple predictor set. To further strengthen model transferability, future research should focus on improving the representativeness of the data. The current study is based on a limited number of European sites. As a result, the dataset does not fully capture variability within ecosystem types, since a single site cannot capture or represent the full range of conditions within a biome, where factors such as community structure or stand age may vary widely. Incorporating additional sites and longer time series will be essential to characterize variability better and increase model robustness. As longer continuous datasets of both in-situ and satellite data become available, future works should also explore temporal validation using independent years or datasets to better assess predictive models.

Expanding the framework to a broader range of sites, ecosystem types, and geographical regions beyond Europe will be necessary not only to increase the robustness of the models but also to strengthen their generalization and assess transferability. Yet, expanding the dataset requires careful consideration, as in global repositories, some ecosystems are better represented than others. Future research should therefore aim to strike a balance between the number of sites, the representativeness of different ecosystem types, and the availability of comparable data lengths across sites, ensuring both diversity and consistency in the training dataset.

The current framework was designed using eddy covariance observations, as EC-derived measurements are essential for training and validating the model. When trained on a more representative dataset spanning a broader range of ecosystems, the framework may support application at new sites without site-specific recalibration. This requires that the same predictor variables are available, whether derived from in-situ observations, satellite or model products. In this context, integrating sensible and latent heat fluxes from model results based on satellite observations (Martens et al., 2020; Miralles et al., 2025) may support broader spatial applicability of the framework. This may also be relevant for future studies aiming to upscale site-level GPP estimates from flux towers (Running et al., 2004; Turner et al., 2006; Jung et al., 2019). Yet, differences in uncertainty, spatial representativeness, and temporal resolution across predictor sources would require careful validation before upscaling.

4.3. Variable contribution

Feature importance provides insights into which indicators were most relevant for predicting GPP in each ecosystem. Across sites, the contribution of remote sensing VIs and in-situ variables varied, reflecting how productivity is regulated differently across ecosystems and how the joint use of in-situ variables (including meteorological drivers and derived flux products) and remote sensing–derived products can provide complementary information on ecosystem functioning.

For the cropland (Estrees Mons) and deciduous broadleaf forest (Hohes Holz), the latent heat turbulent flux was the most influential predictor in the XGBoost model. The latent heat turbulent flux results from the correlation between the measured turbulent flux of water vapor and the measured vertical wind speed, as measured by the eddy-covariance technique. While negative values reflect condensation (or dew) on the leaves, positive values indicate evaporation from water, soil or leaf surfaces, or transpiration from leaves. These processes are driven by the water vapor pressure of the surrounding air and the movement of water through the plant, from the roots, through the leaves and their stomata, to the atmosphere, and are tightly coupled to the photosynthesis rate (He et al., 2022).

For the evergreen needleleaf forest (Hyytiala), air temperature was identified as the most important feature. This finding aligns with previous research. Wu et al. (2012), found that air temperature was the major limiting factor for photosynthesis in early spring, autumn, and winter. This indicates the significant impact of temperature on photosynthetic rates during seasons characterized by lower temperatures, common in the Nordic climate of Hyytiala, and in evergreen needleleaf forests (Chen et al., 2022).

For the grassland (Torgnon), the Red-edge Chlorophyll Index scored as the most important variable. This is consistent with the results of Lin et al. (2019), who found that Clr-based GPP estimates exhibited the highest correlation and low uncertainties with GPP from EC across grassland sites. The occurrence of MDVI, EVI and NDVI as feature importance aligns with the study of Wang et al. (2023), who identified these indices as key predictors of grassland productivity.

The unified models built with XGBoost and LSTM did not benefit from the inclusion of global variables (e.g., elevation, latitude, and longitude) (Tables S6 and S8, supplementary). This result indicates that these factors were either adequately captured by other variables (e.g., air temperature, radiation) or less critical than expected in influencing GPP prediction in different sites.

5. Conclusion

In this study, we evaluated the feasibility of a unified (cross-sites) modeling approach for estimating GPP across multiple ecosystem types without site-specific recalibration. By combining ICOS in-situ measurements with Sentinel-2 derived vegetation indices, we compared a statistical (SARIMAX), a machine learning (XGBoost), and a deep learning (LSTM) approach. In the site-specific setting, the three methods performed similarly, achieving R^2 up to 0.91. SARIMAX served as a site-specific statistical baseline. It struggled with rapid non-linear GPP changes, including peak predictions. XGBoost tended to underestimate high peaks and overestimate low GPP values, particularly during the off-peak season. LSTM showed stronger performance in capturing non-linear GPP dynamics and productivity peaks. All models produced occasional negative GPP predictions under near-zero conditions, highlighting the need for non-negativity constraints in future research. In the cross-site setting, XGBoost and LSTM were employed under the incremental cross-site framework, with XGBoost reaching higher performance than LSTM both on seen (average $R^2 = 0.74$, range 0.65–0.86) and unseen sites (average $R^2 = 0.72$, range 0.66–0.78). Though both models exhibited training instability, XGBoost and LSTM proved to be suitable tools for cross-site GPP prediction.

Overall, this work provides a proof of concept that data-driven frameworks can support spatially transferable GPP monitoring. While the present analysis is based on a limited number of European sites and ecosystem conditions, the results highlight the potential of this approach for cross-site application. Future work should expand the model framework by including additional sites, ecosystem types, ecosystem-specific information, and validating them across a broader geographical range to enhance their robustness, generalization and reliability, especially in global contexts beyond Europe. Further scaling up the modeling framework to generate products at a broader spatial scale and high frequency represents an important next step of this research. Taking this step in this direction, we move closer to improving environmental monitoring and supporting informed ecosystem management using cutting-edge technologies and their developments.

CRediT authorship contribution statement

Anna Spinosa: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Funding acquisition, Conceptualization. **Karisma Karisma:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Marieke A. Eleveld:** Writing – review & editing, Supervision, Conceptualization. **Mario Alberto Fuentes-Monjaraz:** Writing – review & editing, Supervision, Software, Data curation. **Valeria Mobilia:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Ulf Mallast:** Writing – review & editing, Funding acquisition. **Johannes Peterseil:** Writing – review & editing, Funding acquisition. **Ghada El Serafy:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Professor A. Johannes Schmidt-Hieber for his invaluable insights and guidance, Jana Lim, for her support with data collection. Further, we would like to thank the ICOS Research Infrastructure for providing the flux tower data on used in this study. We thank the teams at the ICOS stations of Hyytiala, Torgnon,

Estrees-Mons, Hohes Holz, Klingenberg, Svartberget, Grillenburg and Fontainebleau-Barbeau, for their effort in collecting data. We acknowledge the use of modified Copernicus Sentinel data, 2025. This work was supported by the AGAME project funded by the European Space Agency, France (ESA, contract no. 4000143740/24/I-AG) in the frame of the GEOSS Platform Plus project (Horizon Europe, GA No. GA.Nr. 101039118). The work done is based on the requirements defined in eLTER, also contributing to the development of the eLTER Site Information Cluster.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2026.103820.

Data availability

In-situ data for the model calibration and validation were downloaded from the ICOS Carbon portal (https://doi.org/10.18160/S6HM-CP8Q). The python codes used for the analysis are available at 10.4121/b26f4168-6359-4257-8ef2-3362d6bc6593.

References

Alzubaidi, L., Zhang, J., Humaidi, A., et al., 2021. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J. Big Data 8, 53. http://dx.doi.org/10.1186/s40537-021-00444-8.

Ariyo, A.A., Adewumi, A.O., Ayo, C.K., 2014. Stock price prediction using the arima model. In: 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. IEEE, pp. 106–112. http://dx.doi.org/10.1109/UKSim.2014.67.

Astola, H., Häme, T., Sirro, L., et al., 2019. Comparison of sentinel-2 and landsat 8 imagery for forest variable prediction in boreal region. Remote Sens. Environ. 223, 257–273. http://dx.doi.org/10.1016/j.rse.2019.01.019.

Bernhofer, C., Eichelmann, U., Grünwald, T., Hehn, M., Mauder, M., Moderow, U., Prasse, H., 2025a. ETC L2 ARCHIVE from Klingenberg, 2018–2024. Ecosystem Thematic Centre, URL: https://hdl.handle.net/11676/c8lbCxWNrXNAFQcdJ7aYeHkv.

Bernhofer, C., Eichelmann, U., Grünwald, T., Hehn, M., Mauder, M., Moderow, U., Prasse, H., 2025b. ETC L2 Fluxes from Grillenburg, 2016-12-31–2024-12-31. Ecosystem Thematic Centre, URL: https://hdl.handle.net/11676/vdrkLkMNY5MXqW5htw5adeyw.

Berveiller, D., Dufrêne, E., Delpierre, N., Morfin, A., Francois, C., Vincent, G., Bazot, S., Soudani, K., Girardin, C., Guillot, T., Perot-Guillaume, C., 2025. ETC L2 Fluxes from Fontainebleau-Barbeau, 2018-12-31–2025-09-30. Ecosystem Thematic Centre, URL: https://hdl.handle.net/11676/WgPSFQlVYui2ANucafunTOY9.

Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Cai, Z., Junttila, S., Holst, J., et al., 2021. Modelling Daily Gross Primary Productivity with Sentinel-2 Data in the Nordic Region–Comparison with Data from MODIS. Remote. Sens. 13, URL: https://www.mdpi.com/2072-4292/13/3/469.

Chang, X., Xing, Y., Gong, W., et al., 2023. Evaluating gross primary productivity over 9 ChinaFlux sites based on random forest regression models, remote sensing, and eddy covariance data. Sci. Total Environ. 875, 162601. http://dx.doi.org/10.1016/j.scitotenv.2023.162601, URL: https://www.sciencedirect.com/science/article/pii/S0048969723012172.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. http://dx.doi.org/10.1145/2939672.2939785.

Chen, S., Liu, L., Sui, L., Liu, X., Ma, Y., 2025. An improved spatially downscaled solar-induced chlorophyll fluorescence dataset from the TROPOMI product. Sci. Data 12, 135. http://dx.doi.org/10.1038/s41597-024-04325-6.

Chen, Y., Xu, X., Huang, C., et al., 2022. Selection of prediction factors of gross primary productivity based on artificial neural network. In: 2022 International Conference on Artificial Intelligence, Information Processing and Cloud Computing. AIIPCC, pp. 426–429. http://dx.doi.org/10.1109/AIIPCC57291.2022.00096.

Chicco, D., Warrens, M., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, e623. http://dx.doi.org/10.7717/peerj-cs.623.

Copernicus Sentinel Hub, 2025. Sentinel-2 level-2A data documentation, copernicus sentinel hub documentation. URL: https://documentation.dataspace.copernicus.eu/APIs/SentinelHub/Data/S2L2A.html. (Accessed 19 March 2025).

Cremonese, E., Galvagno, M., di Cella, U. Morra, 2025. ETC L2 Fluxes from Torgnon, 2015-12-31–2024-12-31. Ecosystem Thematic Centre, URL: https://hdl.handle.net/11676/oWX79D_bh4t41D7IWxkmEmsF.

Cui, Z., Qing, X., Chai, H., Yang, S., Zhu, Y., Wang, F., 2021. Real-time rainfall-runoff prediction using light gradient boosting machine coupled with singular spectrum analysis. J. Hydrol. 603, 127124. http://dx.doi.org/10.1016/j.jhydrol.2021.127124.

Das, N., Entekhabi, D., Dunbar, R., Kim, S., Yueh, S., Colliander, A., O'Neill, P., Jackson, T., Jagdhuber, T., Chen, F., et al., 2018. SMAP/Sentinel-1 L2 radiometer/radar 30-second scene 3 km EASE-grid soil moisture, version 2. NASA Natl. Snow Ice Data Cent. Distrib. Act. Arch. Cent. (DAAC) Boulder Color. USA http://dx.doi.org/10.5067/KE1CSVXMI95Y.

Du, D., Zheng, C., Jia, L., Chen, Q., Jiang, M., Hu, G., Lu, J., 2022. Estimation of global cropland gross primary production from satellite observations by integrating water availability variable in light-use-efficiency model. Remote. Sens. 14, 1722. http://dx.doi.org/10.3390/rs14071722.

Ede, S., Baghdadlian, S., Weber, L., et al., 2022. Explain to not forget: Defending against catastrophic forgetting with XAI. URL: https://arxiv.org/abs/2205.01929, arXiv:2205.01929.

ESA, 2025. Sentinel-2 user handbook. https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook.pdf. (Accessed 19 March 2025).

Fathi, M.M., Awadallah, A.G., Abdelbaki, A.M., et al., 2019. A new budyko framework extension using time series SARIMAX model. J. Hydrol. 570, 827–838. http://dx.doi.org/10.1016/j.jhydrol.2019.01.037, URL: https://www.sciencedirect.com/science/article/pii/S0022169419301040.

Fattah, J., Ezzine, L., Aman, Z., et al., 2018. Forecasting of demand using ARIMA model. Int. J. Eng. Bus. Manag. 10, http://dx.doi.org/10.1177/1847979018808673.

Frampton, W.J., Dash, J., Watmough, G., Milton, E.J., 2013. Evaluating the capabilities of sentinel-2 for quantitative estimation of biophysical variables in vegetation. ISPRS J. Photogramm. Remote Sens. 82, 83–92. http://dx.doi.org/10.1016/j.isprsjprs.2013.04.007.

Galvagno, M., Wohlfahrt, G., Cremonese, E., Rossini, M., Colombo, R., Filippa, G., Julitta, T., Manca, G., Siniscalco, C., Cella, U.M. Di, et al., 2013. Phenology and carbon dioxide source/sink strength of a subalpine grassland in response to an exceptionally short snow season. Environ. Res. Lett. 8, 025008. http://dx.doi.org/10.1088/1748-9326/8/2/025008.

Grace, J., 2004. Understanding and managing the global carbon cycle. J. Ecol. 92, 189–202. http://dx.doi.org/10.1111/j.0022-0477.2004.00874.x.

Guo, Y., Lai, X., Gan, M., 2023. Cyanobacterial biomass prediction in a shallow lake using the time series SARIMAX models. Ecol. Informatics 78, 102292, URL: https://www.sciencedirect.com/science/article/pii/S1574954123003217.

He, S., Zhang, Y., Ma, N., Tian, J., Kong, D., Liu, C., 2022. A daily and 500 m coupled evapotranspiration and gross primary production product across China during 2000–2020. Earth Syst. Sci. Data Discuss. 2022, 1–42. http://dx.doi.org/10.5194/essd-14-5463-2022.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hu, C., Hu, S., Zeng, L., et al., 2024. Estimation of daily maize gross primary productivity by considering specific leaf nitrogen and phenology via machine learning methods. Remote. Sens. 16, URL: https://www.mdpi.com/2072-4292/16/2/341.

Integrated Carbon Observation System, 2022. FLUXES - the European greenhouse gas bulletin. https://www.icos-cp.eu/fluxes/1. (Accessed 19 March 2025).

Integrated Carbon Observation System, 2025. ICOS community portal. URL: https://www.icos-cp.eu/. (Accessed 19 March 2025).

Jung, H., Ju, J., Jung, M., et al., 2016. Less-forgetting learning in deep neural networks. ArXiv abs/1607.00122, URL: https://api.semanticscholar.org/CorpusID:18398195.

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., Reichstein, M., 2019. The FLUXCOM ensemble of global land–atmosphere energy fluxes. Sci. Data 6, 74. http://dx.doi.org/10.1038/s41597-019-0076-8.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30, URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Khatib, A.E., Karray, F., 2019. Preempting catastrophic forgetting in continual learning models by anticipatory regularization. In: 2019 International Joint Conference on Neural Networks. IJCNN, pp. 1–7. http://dx.doi.org/10.1109/IJCNN.2019.8852426.

Kumar, N., Jain, V., Joshi, K., et al., 2023. Prediction of epidemic disease cases using ARIMA and SARIMAX models. In: 2023 Sixth International Conference of Women in Data Science At Prince Sultan University. WiDS PSU, pp. 201–205. http://dx.doi.org/10.1109/WiDS-PSU57071.2023.00049.

Lai, J., Zhang, Y., Wang, A., Fei, W., Diao, Y., Li, R., Wu, J., 2025. FLAML version 2.3.3: Model-based assessment of gross primary productivity at forest, grassland, and cropland ecosystem sites. Geosci. Model. Dev. Discuss. 2025, 1–54. http://dx.doi.org/10.5194/gmd-18-5115-2025.

Lasslop, G., Reichstein, M., Papale, D., Richardson, A.D., Arneth, A., Barr, A., Stoy, P., Wohlfahrt, G., 2010. Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. Global Change Biol. 16, 187–208. http://dx.doi.org/10.1111/j.1365-2486.2009.02041.x.

Lee, B., Kim, N., Kim, E.-S., Jang, K., Kang, M., Lim, J.-H., Cho, J., Lee, Y., 2020. An artificial intelligence approach to predict gross primary productivity in the forests of South Korea using satellite remote sensing data. Forests 11, 1000. http://dx.doi.org/10.3390/f11091000.

Leonard, J., Bornet, F., François, B., Grehan, E., 2025. ETC L2 Fluxes from Estrees-Mons A28, 2016-12-31–2024-12-31. Ecosystem Thematic Centre, URL: https://hdl.handle.net/11676/4lrU6O5qv9KCuvuhMM6veY8D.

Liao, Z., Zhou, B., Zhu, J., Jia, H., Fei, X., 2023. A critical review of methods, principles and progress for estimating the gross primary productivity of terrestrial ecosystems. Front. Environ. Sci. 11, 1093095.

Lin, S., Li, J., Liu, Q., et al., 2019. Evaluating the effectiveness of using vegetation indices based on red-edge reflectance from sentinel-2 to estimate gross primary productivity. Remote. Sens. 11, http://dx.doi.org/10.3390/rs11111303, URL: https://www.mdpi.com/2072-4292/11/11/1303.

Liu, S., Zhuang, Q., He, Y., Noormets, A., Chen, J., Gu, L., 2016. Evaluating atmospheric CO2 effects on gross primary productivity and net ecosystem exchanges of terrestrial ecosystems in the conterminous united states using the ameriflux data and an artificial neural network approach. Agricult. Forest. Meterol. 220, 38–49. http://dx.doi.org/10.1016/j.agrformet.2016.01.007.

Liu, J., Zuo, Y., Wang, N., Yuan, F., Zhu, X., Zhang, L., Zhang, J., Sun, Y., Guo, Z., Guo, Y., 2021. Comparative analysis of two machine learning algorithms in predicting site-level net ecosystem exchange in major biomes. Remote. Sens. 13, 2242. http://dx.doi.org/10.3390/rs13122242.

LSTM layer, Keras Developers, 2025. LSTM layer - keras documentation. https://keras.io/api/layers/recurrent_layers/lstm/. (Accessed 19 March 2025).

Lu, Q., Liu, H., Wei, L., Zhong, Y., Zhou, Z., 2024. Global prediction of gross primary productivity under future climate change. Sci. Total Environ. 912, 169239. http://dx.doi.org/10.1016/j.scitotenv.2023.169239.

Mammarella, I., Aalto, J., Back, J., Kolari, P., Laakso, H., Levula, J., Matilainen, T., Pihlatie, M., Pumpanen, J., Taipale, R., Vesala, T., 2023. ETC L2 Fluxes from Hyytiala, 2017-12-31–2023-10-31. Ecosystem Thematic Centre, URL: https://hdl.handle.net/11676/Ke4Y-W71Kj_Tc8l2Q–ZZpMZ.

Martens, B., Schumacher, D.L., Wouters, H., Muñoz-Sabater, J., Verhoest, N.E.C., Miralles, D.G., 2020. Evaluating the land-surface energy partitioning in ERA5. Geosci. Model. Dev. 13, 4159–4181. http://dx.doi.org/10.5194/gmd-13-4159-2020.

Miralles, D.G., Bonte, O., Koppa, A., Baez-Villanueva, O.M., Tronquo, E., Zhong, F., Beck, H.E., Hulsman, P., Dorigo, W., Verhoest, N.E.C., et al., 2025. GLEAM4: Global land evaporation and soil moisture dataset at 0.1° resolution from 1980 to near present. Sci. Data 12, 416. http://dx.doi.org/10.1038/s41597-025-04747-1.

Montero, D., Mahecha, M.D., Martinuzzi, F., Aybar, C., Klosterhalfen, A., Knohl, A., Koebsch, F., Anaya, J., Wieneke, S., 2024. Recurrent neural networks for modelling gross primary production. In: IGARSS 2024–IEEE International Geoscience and Remote Sensing Symposium. pp. 4214–4217. http://dx.doi.org/10.1109/IGARSS53475.2024.10640715.

Na, Q., Lai, Q., Bao, G., Xue, J., Liu, X., Gao, R., 2025. Estimation of gross primary productivity using performance-optimized machine learning methods for the forest ecosystems in China. Forests 16, 518. http://dx.doi.org/10.3390/f16030518.

Noumonvi, K., Ferlan, M., Eler, K., et al., 2019. Estimation of carbon fluxes from eddy covariance data and satellite-derived vegetation indices in a karst grassland (podgorski kras, Slovenia). Remote. Sens. 11, 649. http://dx.doi.org/10.3390/rs11060649.

Owen, K.E., Tenhunen, J., Reichstein, M., Wang, Q., Falge, E., Geyer, R., Xiao, X., Stoy, P., Ammann, C., Arain, A., et al., 2007. Linking flux network measurements to continental scale simulations: Ecosystem carbon dioxide exchange capacity under non-water-stressed conditions. Global Change Biol. 13, 734–760. http://dx.doi.org/10.1111/j.1365-2486.2007.01326.x.

Papale, D., Reichstein, M., Aubinet, M., et al., 2006. Towards a standardized processing of net ecosystem exchange measured with eddy covariance technique: Algorithms and uncertainty estimation. URL: www.biogeosciences.net/3/571/2006/.

Pastorello, G., Trotta, C., Canfora, E., et al., 2020. The FLUXNET2015 dataset and the oneflux processing pipeline for eddy covariance data. Sci. Data 7, 1–27. http://dx.doi.org/10.1038/s41597-020-0534-3, URL: https://www.nature.com/articles/s41597-020-0534-3.

Peichl, M., Nilsson, M., Larmanou, E., Smith, P., Marklund, P., Simon, G. De, Lofvenius, P., Dignam, R., Holst, J., Molder, M., Andersson, T., Boschetti, F., Kozii, N., Linderson, M.-L., Ottosson-Löfvenius, M., 2025. ETC L2 Fluxnet (half-hourly) from Svartberget, 2018-12-31–2025-09-30. Ecosystem Thematic Centre, URL: https://hdl.handle.net/11676/qR-2VNMiMXAsYwMLzwucjOsR.

Pluntke, T., Bernhofer, C., Grünwald, T., Renner, M., Prasse, H., 2023. Long-term climatological and ecohydrological analysis of a paired catchment – flux tower observatory near dresden (germany). is there evidence of climate change in local evapotranspiration? J. Hydrol. 617, 128873, URL: https://www.sciencedirect.com/science/article/pii/S0022169422014433.

Purdy, A.J., Fisher, J.B., Goulden, M.L., Colliander, A., Halverson, G., Tu, K., Famiglietti, J.S., 2018. SMAP soil moisture improves global evapotranspiration. Remote Sens. Environ. 219, 1–14. http://dx.doi.org/10.1016/j.rse.2018.09.023.

Rebmann, C., Dienstbach, L., Schmidt, P., Wiesen, R., Meis, J., Feldmann, I., Campos, F. Bastos, Dejosez, S., Quiros, I. Garcia, Gimper, S., Hautmann, D., Hildebrandt, A., Kempka, P., Paasch, S., 2025. ETC L2 Fluxes from Hohes Holz, 2018-12-31–2025-09-30. Ecosystem Thematic Centre, URL: https://hdl.handle.net/11676/2L1kNpUSC5efdHkzpoNCpyDT.

Robinson, N.P., Allred, B.W., Smith, W.K., Jones, M.O., Moreno, A., Erickson, T.A., Naugle, D.E., Running, S.W., 2018. Terrestrial primary production for the conterminous united states derived from landsat 30 m and modis 250 m. Remote. Sens. Ecol. Conserv. 4, 264–280. http://dx.doi.org/10.1002/rse2.74.

Rossini, M., Migliavacca, M., Galvagno, M., Meroni, M., Cogliati, S., Cremonese, E., Fava, F., Gitelson, A., Julitta, T., di Cella, U.M., et al., 2014. Remote estimation of grassland gross primary production during extreme meteorological seasons. Int. J. Appl. Earth Obs. Geoinf. 29, 1–10. http://dx.doi.org/10.1016/j.jag.2013.12.008.

Rubio-Cuadrado, Á., Gómez, C., Rodríguez-Calcerrada, J., Perea, R., Gordaliza, G.G., Camarero, J.J., Montes, F., Gil, L., 2021. Differential response of oak and beech to late frost damage: an integrated analysis from organ to forest. Agricult. Forest. Meterol. 297, 108243. http://dx.doi.org/10.1016/j.agrformet.2020.108243.

Running, S.W., Nemani, R.R., Heinsch, F.A., Zhao, M., Reeves, M., Hashimoto, H., 2004. A continuous satellite-derived measure of global terrestrial primary production. BioScience 54, 547–560. http://dx.doi.org/10.1641/0006-3568(2004)054[0547:ACSMOG]2.0.CO;2.

SARIMAX, Statsmodels Developers, 2025. SARIMAX — statsmodels 0.13.2 documentation. https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html. (Accessed 19 March 2025).

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639. http://dx.doi.org/10.1021/ac60214a047, publisher: American Chemical Society.

Seabold, S., Perktold, J., 2010. Statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference.

Spinosa, A., Fuentes-Monjaraz, M.A., Serafy, G. El, 2023. Assessing the use of sentinel-2 data for spatio-temporal upscaling of flux tower gross primary productivity measurements. Remote. Sens. 15, 562. http://dx.doi.org/10.3390/rs15030562.

Tiwari, A., 2022. Chapter 2 - supervised learning: From theory to applications. In: Pandey, R., Khatri, S.K., kumar Singh, N., Verma, P. (Eds.), Artificial Intelligence and Machine Learning for EDGE Computing. Academic Press, pp. 23–32, URL: https://www.sciencedirect.com/science/article/pii/B9780128240540000265.

Tolcha, T.D., 2023. The state of africa's air transport market amid COVID-19, and forecasts for recovery. J. Air Transp. Manag. 108, 102380, URL: https://www.sciencedirect.com/science/article/pii/S0969699723000236.

Tramontana, G., Ichii, K., Camps-Valls, G., Tomelleri, E., Papale, D., 2015. Uncertainty analysis of gross primary production upscaling using random forests, remote sensing and eddy covariance data. Remote Sens. Environ. 168, 360–373. http://dx.doi.org/10.1016/j.rse.2015.07.015.

Turner, D.P., Ritts, W.D., Cohen, W.B., Gower, S.T., Running, S.W., Zhao, M., Costa, M.H., Kirschbaum, A.A., Ham, J.M., Saleska, S.R., et al., 2006. Evaluation of MODIS NPP and GPP products across multiple biomes. Remote Sens. Environ. 102, 282–292. http://dx.doi.org/10.1016/j.rse.2006.02.017.

Wang, Y., Li, R., Hu, J., et al., 2021. Daily estimation of gross primary production under all sky using a light use efficiency model coupled with satellite passive microwave measurements. Remote Sens. Environ. 267, 112721, URL: https://www.sciencedirect.com/science/article/pii/S0034425721004417.

Wang, H., Shao, W., Hu, Y., et al., 2023. Assessment of Six Machine Learning Methods for Predicting Gross Primary Productivity in Grassland. Remote. Sens. 15, http://dx.doi.org/10.3390/rs15143475, publisher: Multidisciplinary Digital Publishing Institute (MDPI).

Wu, S.H., Jansson, P.-E., Kolari, P., 2012. The role of air and soil temperature in the seasonality of photosynthesis and transpiration in a boreal scots pine ecosystem. Agricult. Forest. Meterol. 156, 85–103. http://dx.doi.org/10.1016/j.agrformet.2012.01.006, URL: https://www.sciencedirect.com/science/article/pii/S0168192312000226.

XGBoost, XGBoost Developers, 2025a. Xgboost parameters. https://xgboost.readthedocs.io/en/stable/parameter.html. (Accessed 19 March 2025).

XGBoost, XGBoost developers, 2025b. XGBoost python package. https://xgboost.readthedocs.io/en/stable/python/index.html. (Accessed 19 March 2025).

Xing, X., Wu, M., Zhu, H., Duan, W., Ju, W., Wang, X., Ran, Y., Zhang, Y., Jiang, F., 2025. Optimized gross primary productivity over the croplands within the beps particle filtering data assimilation system (beps_pf v1. 0). J. Adv. Model. Earth Syst. 17, e2024MS004412. http://dx.doi.org/10.1029/2024MS004412.

Yuan, W., Cai, W., Nguy-Robertson, A.L., Fang, H., Suyker, A.E., Chen, Y., Dong, W., Liu, S., Zhang, H., 2015. Uncertainty in simulating gross primary production of cropland ecosystem from satellite-based models. Agricult. Forest. Meterol. 207, 48–57. http://dx.doi.org/10.1016/j.agrformet.2015.03.016.

Zhang, T., Zhou, J., Yu, P., et al., 2023. Response of ecosystem gross primary productivity to drought in northern China based on multi-source remote sensing data. J. Hydrol. 616, 128808. http://dx.doi.org/10.1016/j.jhydrol.2022.128808, URL: https://www.sciencedirect.com/science/article/pii/S0022169422013786.

Zhao, X., Zhao, P., Zhu, L., Zhang, G., 2022. A comparison of multivariate and univariate time series models applied in tree sap flux analyses. For. Sci. 68, 473–486. http://dx.doi.org/10.1093/forsci/fxac027.

Zhu, W., Xie, Z., Zhao, C., Zheng, Z., Qiao, K., Peng, D., Fu, Y.H., 2024. Remote sensing of terrestrial gross primary productivity: a review of advances in theoretical foundation, key parameters and methods. GIScience Remote. Sens. 61, 2318846. http://dx.doi.org/10.1080/15481603.2024.2318846.