Title
Using automated machine learning for the upscaling of gross primary productivity

Journal
Biogeosciences, 21(10)

ISSN
1726-4170

Authors
Gaber, Max
Kang, Yanghui
Schurgers, Guy
et al.

Publication Date
2024

DOI
10.5194/bg-21-2447-2024

Biogeosciences, 21, 2447–2472, 2024
https://doi.org/10.5194/bg-21-2447-2024

Biogeosciences Open Access EGU

Using automated machine learning for the upscaling of gross primary productivity

Max Gaber[1,2], **Yanghui Kang**[1,3], **Guy Schurgers**[2], **and Trevor Keenan**[1,3]

[1]Department of Environmental Science, Policy, and Management, UC Berkeley, Berkeley, CA 94720, USA
[2]Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, 1350, Denmark
[3]Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Correspondence: Max Gaber (mfg@ign.ku.dk), Yanghui Kang (yanghuikang@berkeley.edu), and Trevor Keenan (trevorkeenan@berkeley.edu)

Received: 25 August 2023 – Discussion started: 31 August 2023
Revised: 23 March 2024 – Accepted: 7 April 2024 – Published: 24 May 2024

Abstract. Estimating gross primary productivity (GPP) over space and time is fundamental for understanding the response of the terrestrial biosphere to climate change. Eddy covariance flux towers provide in situ estimates of GPP at the ecosystem scale, but their sparse geographical distribution limits larger-scale inference. Machine learning (ML) techniques have been used to address this problem by extrapolating local GPP measurements over space using satellite remote sensing data. However, the accuracy of the regression model can be affected by uncertainties introduced by model selection, parameterization, and choice of explanatory features, among others. Recent advances in automated ML (AutoML) provide a novel automated way to select and synthesize different ML models. In this work, we explore the potential of AutoML by training three major AutoML frameworks on eddy covariance measurements of GPP at 243 globally distributed sites. We compared their ability to predict GPP and its spatial and temporal variability based on different sets of remote sensing explanatory variables. Explanatory variables from only Moderate Resolution Imaging Spectroradiometer (MODIS) surface reflectance data and photosynthetically active radiation explained over 70 % of the monthly variability in GPP, while satellite-derived proxies for canopy structure, photosynthetic activity, environmental stressors, and meteorological variables from reanalysis (ERA5-Land) further improved the frameworks' predictive ability. We found that the AutoML framework Auto-sklearn consistently outperformed other AutoML frameworks as well as a classical random forest regressor in predicting GPP but with small performance differences, reaching an r^2 of up to 0.75. We deployed the best-performing framework to generate global wall-to-wall maps highlighting GPP patterns in good agreement with satellite-derived reference data. This research benchmarks the application of AutoML in GPP estimation and assesses its potential and limitations in quantifying global photosynthetic activity.

1 Introduction

Terrestrial gross primary productivity (GPP) describes the gross photosynthetic assimilation of atmospheric carbon dioxide (CO_2) at the ecosystem scale. As the largest flux in the global carbon cycle, GPP plays a vital role in maintaining ecosystem functions and sustaining human well-being (Beer et al., 2010; Friedlingstein et al., 2019). In addition, the dynamics of GPP directly affect the growth rate of atmospheric CO_2 concentrations and ecosystem feedbacks to the climate system. Therefore, accurate estimates of the magnitude and spatiotemporal patterns of terrestrial GPP are essential for understanding ecosystem carbon cycling and developing effective climate change mitigation and adaptation strategies (Keenan et al., 2016; Canadell et al., 2021).

While in situ GPP estimates are available from methods such as the eddy covariance technique, global spatiotemporal patterns are challenging to estimate due to the lack of large-scale observations and the high uncertainty of process-based vegetation models (Anav et al., 2015). Fluxes captured by the eddy covariance measurements are limited to the area within the tower's footprint, typically ranging from several hun-

dred meters to several kilometers (Gong et al., 2009). Therefore, various data-driven methods such as machine learning (ML) have been used to scale up in situ GPP measurements from flux tower networks to a global scale. These ML models use independent globally available explanatory data from remote sensing or other continuous model outputs to infer a functional relationship to the GPP measurements, which can be used to predict GPP in areas beyond the limited flux tower footprints. Commonly applied models include tree-based methods (Bodesheim et al., 2018; Wei et al., 2017; Beer et al., 2010; Jung et al., 2011), artificial neural networks (Joiner and Yoshida, 2020; Beer et al., 2010; Papale et al., 2015), linear regressors, kernel methods, and ensembles thereof (Tramontana et al., 2016). Despite the wide variety of ML models applied, a high degree of uncertainty remains in the selection of appropriate features, algorithms, and configurations (Reichstein et al., 2019). The data-based models typically perform well in estimating seasonal GPP patterns but show limitations in predicting trends and interannual variability (Tramontana et al., 2016).

The contribution of different explanatory variables, such as greenness measures, photosynthetically active radiation (PAR), land surface temperature (LST), soil moisture (SM), and meteorological variables (vapor pressure deficit, temperature, precipitation) to the accuracy of the GPP predictions (hereafter referred to as variable importance) has not been conclusively clarified. Both Tramontana et al. (2016) and Joiner and Yoshida (2020) confirmed the dominant control of remotely sensed greenness on the ML prediction of GPP at daily to interannual timescales, with meteorological variables contributing marginally. Conversely, Stocker et al. (2018) found an important control of site-measured soil moisture on light use efficiency (LUE) and GPP at daily granularity under drought conditions at flux sites. Furthermore, Dannenberg et al. (2023) showed that including satellite-derived soil moisture and LST data significantly improved the estimation of monthly GPP in drylands over the western US. However, a comprehensive assessment of the importance of meteorological and satellite-derived variables beyond vegetation structure at the global scale is lacking. Given the ubiquitous intercorrelation between remote sensing and meteorological variables, the importance of different explanatory variables has typically been accomplished by training separate models on different input combinations (Tramontana et al., 2016). Yet, ML model performance can vary strongly depending on the dimension of input features, hyperparameter tuning (the search for the optimal parameters that control the learning process of an ML model), and even the specific type of ML model employed (Raschka, 2020; Cawley and Talbot, 2010). Therefore, a unified ML framework that concurrently optimizes model choice and parameterization is required to facilitate a balanced assessment of driver importance in global GPP upscaling.

Navigating the search space created by the choice of model architecture, hyperparameters, and preprocessing steps to find a suitable combination for GPP prediction is a resource-intensive task. Therefore, researchers often evaluate a selection of combinations that they expect to perform well, thereby potentially missing out on the optimal solution (Karmaker et al., 2021). Automated machine learning (AutoML) aims to overcome these challenges through an autonomous approach. By evaluating different combinations of preprocessing steps, candidate ML models, and hyperparameters, AutoML aims to find the optimal ML configuration for the given ML problem and available training data. In addition, it leverages the unique strengths of different algorithms by using ensembling or stacking techniques. At the time of this study, AutoML was still under ongoing development but has recently received increasing attention in the environmental sciences and beyond. It has shown superior performance to classical ML, for example, in modeling water nutrient concentrations (Kim et al., 2020), dam water inflows (Lee et al., 2023), and water quality prediction (Madni et al., 2023), and similar performance to reference models for climate zone classification (Traoré et al., 2021) and drought forecasts (Duan and Zhang, 2022). Other use cases include predicting landslide hazards (Qi et al., 2021), root zone soil moisture (Babaeian et al., 2021), or GPP at a single flux tower site (Guevara-Escobar et al., 2021).

In this study, we investigate if and how AutoML can improve global GPP upscaling at the monthly frequency from in situ measurements using globally available explanatory variables. We examine the three frameworks Auto-sklearn, H2O AutoML, and AutoGluon in this study since they have shown outstanding performance in benchmarks and Kaggle competitions (Guyon et al., 2019; Erickson et al., 2020; Truong et al., 2019; LeDell and Poirier, 2020; Feurer et al., 2018). All frameworks differ in their architecture and approach to selecting ML algorithms. We evaluate their selection of processing and ML algorithms based on site-level measurements. In addition, we evaluate the variable importance, i.e., the contribution of various remotely sensed vegetation structure variables, proxies for photosynthetic activity and environmental stress (i.e., greenness, land surface temperature, soil moisture, evapotranspiration), and meteorological factors, to the performance of the AutoML frameworks. The impacts of the spatial resolution of remote sensing data on GPP estimation are further assessed. Finally, we upscale our results to global wall-to-wall GPP maps and evaluate their spatiotemporal patterns and associated uncertainties.

2 Methods and materials

2.1 Data

2.1.1 Eddy covariance measurements

We merged eddy covariance datasets from FLUXNET 2015 (Pastorello et al., 2020), AmeriFlux FLUXNET

(https://ameriflux.lbl.gov/data/flux-data-products, last access: 9 June 2023), and ICOS Warm Winter 2020 (Warm Winter 2020 Team and ICOS Ecosystem Thematic Centre, 2022) to obtain a large number of monthly GPP estimates from net ecosystem exchange (NEE) measurements. Where sites were available in more than one source, we kept the most recent record. The data quality control followed previous studies (Tramontana et al., 2016; Jung et al., 2011; Joiner et al., 2018). We considered monthly values where at least 80 % of the NEE data came from actual measurements or were high-quality gap filled. We used the GPP derived from NEE using the nighttime partitioning approach (Reichstein et al., 2005), and negative GPP outliers were truncated at $-1\,\mathrm{gC\,m^{-2}\,d^{-1}}$ average daily GPP.

The preprocessing resulted in a dataset of 243 sites and 18 218 site months, ranging from 2001 to 2020, and serving as the ground truth for the evaluation of site-level GPP predictions (Fig. 1). The distribution of sites and site months shows strong biases in region, biome, and climate representation (Fig. 2). We reorganized the land cover classes, as individual land cover classes related to shrublands and savannas rarely occurred. Therefore, "open shrublands" and "closed shrublands" were merged, as well as "savannas" and "woody savannas", resulting in the following land cover according to the International Geosphere–Biosphere Programme (IGBP) (International Geosphere–Biosphere Programme, 2024): croplands (CROs), shrublands (SHs), deciduous broadleaf forests (DBFs), evergreen broadleaf forests (EBFs), evergreen needleleaf forests (ENFs), grasslands (GRAs), mixed forests (MFs), savannas (SAVs), permanent wetlands (WETs), the non-vegetated classes of permanent snow and ice (SNO), water bodies (WATs), and barren soil (BAR).

2.1.2 Explanatory variables

Our goal was to provide as many explanatory variables as possible and let the frameworks decide which to use. We obtained gridded explanatory variables from various sources of remotely sensed and modeled data with global coverage. The data allowed us to evaluate locally by sampling at the tower locations and to predict on a global wall-to-wall scale. These variables include products based on Moderate Resolution Imaging Spectroradiometer (MODIS) measurements, such as nadir-bidirectional reflectance distribution function (BRDF)-adjusted reflectances (NBARs) from optical to infrared wavelengths, the fraction of photosynthetically active radiation (FPAR), leaf area index (LAI), day and night surface temperature, and land cover. We also included the photosynthetically active radiation (PAR), diffuse PAR, and the surface downwelling shortwave flux (RSDN) from BESS_Rad, as well as solar-induced fluorescence (SIF), evapotranspiration (ET), and soil moisture (SM). In addition, we used meteorological data from the ERA5-Land reanalysis, including precipitation, temperature, and vapor pressure deficit (VPD). We applied a 3-month lag in precipitation to account for water availability. Table 1 shows an overview of all explanatory variables.

Many of the explanatory variables are themselves datasets that have been modeled from MODIS data. For instance, SIF was predicted from MODIS NBAR using a feed-forward neural network trained on OCO-2 SIF retrievals (Zhang et al., 2018). ET estimates were modeled by a coupled land sur-

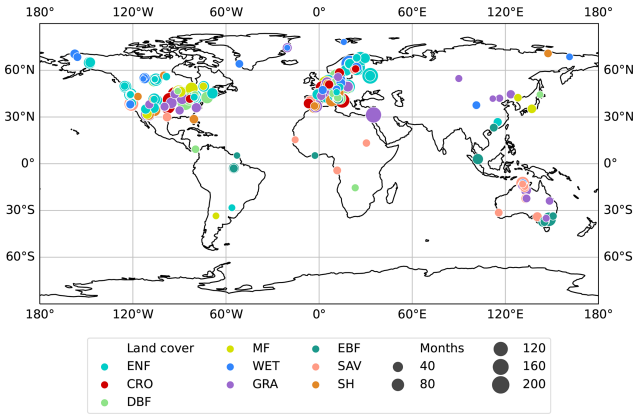

Figure 1. Locations of the measurement sites. The marker size represents the number of monthly measurements available at the respective location. The color stands for the land cover class reported at the site and comprises croplands (CROs), shrublands (SHs), deciduous broadleaf forests (DBFs), evergreen broadleaf forests (EBFs), evergreen needleleaf forests (ENFs), grasslands (GRAs), mixed forests (MFs), savannas (SAVs), and permanent wetlands (WETs).

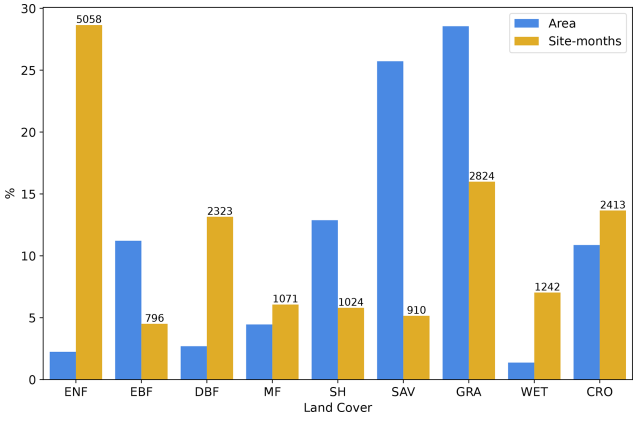

Figure 2. Standardized number of site months and global area of each land cover type, excluding land covers without any GPP measurements. The number of site months is shown above their respective columns. The land cover classes reported follow the IGBP classification (International Geosphere–Biosphere Programme, 2024) and comprise croplands (CROs), shrublands (SHs), deciduous broadleaf forests (DBFs), evergreen broadleaf forests (EBFs), evergreen needleleaf forests (ENFs), grasslands (GRAs), mixed forests (MFs), savannas (SAVs), and permanent wetlands (WETs).

face and atmospheric boundary layer model (Atmosphere–Land Exchange Inverse, ALEXI), which used MODIS LST and LAI as inputs, among others (Hain and Anderson, 2017). Although their input data largely overlap with the inputs to our model, we expected additional improvements from these datasets due to the domain knowledge of their models, which would otherwise be difficult to replicate in this study by solely relying on MODIS data and limited GPP measurements.

We filtered the data for poor-quality pixels, performed gap filling, and matched spatial and temporal resolutions. We used NBAR (MCD43C4 v006), where more than 75 % of high-resolution NBAR pixels were available from the full BRDF inversion. We selected LST data by applying the quality control mask and where the average emissivity error was less than 0.02. LAI and FPAR were used when retrieved using the main algorithm with or without saturation. Data gaps were filled at the native resolution, similar to the procedure of Walther et al. (2022). We filled gaps of less than or equal to 5 d (8 d for 4 d resolution datasets) with the average of a 15 d moving window for high-frequency datasets (NBAR, LAI, FPAR, BESS_Rad, CSIF). We gap filled LST with a 9 d moving window because we observed higher variations. For SM, we used the moving window median for short gaps and the mean seasonal cycle for long gaps. Finally, we resampled all datasets to 0.05° spatial resolution and monthly temporal resolution. Coarser-resolution datasets were resampled using a nearest-neighbor approach, while high-resolution data were downsampled using the conservative remapping method (Jones, 1999).

2.2 Automated machine learning

The performance of ML is highly dependent on the selection and configuration of preprocessing steps, model architectures, and corresponding hyperparameters, which are determined by the specific ML problem (Hutter et al., 2019). The steps involved are typically organized sequentially in an ML pipeline and transform the input features (explanatory variables) into a target variable (Zöller and Huber, 2021). The pipeline refers to the entire process of developing and training an ML model and typically consists of several tasks, such as preprocessing, feature engineering, model training, hyperparameter tuning, and model deployment.

Selecting the appropriate algorithms and hyperparameters is often referred to as the combined algorithm selection and hyperparameter tuning (CASH) problem and involves exploiting a search space spanned by the available algorithms and their parameters. Solving the CASH problem is challenging because the search space is high-dimensional and hierarchical, and its exhaustive exploitation is often computationally expensive (Kotthoff et al., 2019; Thornton et al., 2013). As a result, candidate pipeline configurations are typically determined in controlled experiments using optimization methods, such as grid search, randomized search, and Bayesian optimization, or through experience and educated guesswork (Karmaker et al., 2021).

In contrast, AutoML provides an optimization approach with an end-to-end scope. A fully developed AutoML framework iteratively selects the pipeline structure, algorithms, and hyperparameters from the search space based on data requirements and objective functions while considering a time and resource budget (Yao et al., 2019). Thus, it facilitates usability for domain experts and overcomes inefficient trial-and-error approaches. AutoML draws from a pool of classical ML algorithms (base models) and preprocessing methods and selects or combines the most appropriate candidates for the ML problem. Typically, AutoML frameworks create model ensembles by combining the predictions of their base models, either through a simple aggregation or through yet another model that uses the predictions of the base models as input features. This approach is often superior to individual predictions because it can overcome the limitations of the individual base models (van der Laan et al., 2007).

AutoML frameworks handle pipeline creation with various degrees of autonomy and scope, given the early-stage development of much of the available software at the time of this study. For example, tasks such as pipeline selection or feature engineering are only sporadically implemented in the available frameworks (Zöller and Huber, 2021). With H2O AutoML, Auto-sklearn, and AutoGluon, we compared AutoML frameworks that differ in training procedure, optimization method, and available base models and have been tested in a wide range of applications and benchmarks (Balaji and Allen, 2018; Truong et al., 2019; Erickson et al., 2020; Hanussek et al., 2020; Ferreira et al., 2021).

Auto-sklearn

Auto-sklearn (Feurer et al., 2015a) is an AutoML library built on top of the Scikit-learn ML models. We used Auto-sklearn in version 0.14.7. The framework relies on a wide range of base models, including AdaBoost, ARD (automatic relevance determination) regression, decision trees, extra trees, Gaussian processes, gradient boosting, k-nearest neighbors, support vector regression, MLP (multi-layer perceptron) regression, random forests, and SGD (linear stochastic gradient descent) regression. It also considers feature engineering algorithms, such as PCA (principal component analysis), percentile regression, and feature agglomeration (Feurer et al., 2015b). The framework selects and tunes its base models in a Bayesian optimization and performs a forward stepwise ensemble selection (Caruana et al., 2004). During this process, the framework draws on a pool of ML models to build the model ensemble, but instead of using the entire pool, it adds the models one by one, only using the ones that maximize ensemble performance. Auto-sklearn also uses a meta-learner trained on the meta-features of a variety of datasets to warm start the optimization procedure, which increases efficiency and reduces training time (Feurer

Table 1. Explanatory variables and sources and their respective spatial and temporal resolution.

Explanatory variable	Source	Spatial resolution	Temporal resolution
Reflectance (nadir-BRDF adjusted; NBAR) bands 1–7	MODIS MCD43C4 v006 (Schaaf and Wang, 2015)	0.05°	daily
PAR	BESS_Rad (Ryu et al., 2018)	0.05°	daily
Diffuse PAR	BESS_Rad (Ryu et al., 2018)	0.05°	daily
RSDN	BESS_Rad (Ryu et al., 2018)	0.05°	daily
FPAR	MODIS MCD15A2H v006 (Myneni et al., 2015)	500 m	4 d
LAI	MODIS MCD15A2H v006 (Myneni et al., 2015)	500 m	4 d
Land surface temperature (day)	MODIS MYD11A1, MOD11A1 (Wan et al., 2015)	1 km	daily
Land surface temperature (night)	MODIS MYD11A1, MOD11A1 (Wan et al., 2015)	1 km	daily
Evapotranspiration	ALEXI (Hain and Anderson, 2017)	0.05°	daily
Soil moisture	ESA CCI v.06.1 (Gruber et al., 2019; Dorigo et al., 2017)	0.25°	daily
SIF	CSIF (Zhang et al., 2018)	0.05°	4 d
Instantaneous SIF	CSIF (Zhang et al., 2018)	0.05°	4 d
Land cover (biome)	MODIS MCD12Q1 (Friedl and Sulla-Menashe, 2019)	500 m	annual
Total precipitation	ERA5-Land (Muñoz-Sabater et al., 2021)	0.1°	hourly
Total precipitation (3-month lag)	ERA5-Land (Muñoz-Sabater et al., 2021)	0.1°	hourly
Temperature	ERA5-Land (Muñoz-Sabater et al., 2021)	0.1°	hourly
Vapor pressure deficit (VPD)	ERA5-Land (Muñoz-Sabater et al., 2021)	0.1°	hourly

et al., 2015a). The meta-learner uses knowledge from previous experiments with similar datasets and can, therefore, select promising ML models to start with instead of training from scratch each time.

H2O AutoML

H2O AutoML (LeDell and Poirier, 2020) is a widely used AutoML framework for supervised regression and classification. We used H2O 3 and the Python package of version 3.18.0.2. H2O AutoML draws from a set of base models, which, in the developer's terminology, are divided into the model families of gradient boosting models (GBMs), XGBoost GBMs, generalized linear models (GLMs), a default random forest model (DRF), extremely randomized trees (XRTs), and feed-forward neural networks. The framework trains these models in a predefined order with increasing diversity and complexity, using pre-specified hyperparameters or tuning them by random search. In addition to the individual base models, H2O AutoML creates ensembles of the base models, combining their predictions through a GLM by default. The ensembles consist of either all base models or only the best-performing base models from each model family.

H2O AutoML then ranks the performance of individual models and model ensembles using an internal cross-validation (CV). The best-performing model is used for prediction.

AutoGluon

AutoGluon Tabular (Erickson et al., 2020) relies heavily on ensemble and stacking techniques. It differs from many other frameworks by omitting model selection and hyperparameter tuning, thus avoiding the computationally intensive CASH problem. The framework draws from a pool of base models: neural networks, LightGBM boosted trees, random forests, extremely randomized trees, and k-nearest neighbors. These models are combined in a multi-layer stack ensembling process: AutoGluon first generates predictions from each base model. The predictions are then concatenated with the original features and passed to another set of models (the stacker models) in the next layer. Their predictions can be concatenated again and passed to the next layer, and so on, creating a layered structure of model sets and concatenation steps. The predictions of the last layer are combined in an ensemble selection step (Caruana et al., 2004). Each layer consists of the same base model types and hyperparameters. In addition, AutoGluon implements k-fold bagging, which improves performance by using the training data more efficiently. Global and model-specific preprocessing algorithms are available to impute missing values or correct skewed distributions. A feature selection algorithm is provided in the framework but is still in an experimental stage and not enabled in the version used.

2.3 Experimental design

We first evaluated the three AutoML frameworks under three sets of explanatory variables. In addition, we trained a classical random forest model in a randomized search, which served as our baseline. We then used Auto-sklearn with the best-performing set of explanatory variables to upscale in situ eddy covariance GPP measurements to global wall-to-wall maps (Fig. 3).

2.3.1 Explanatory variable sets

We organized the explanatory variables into three sets to determine their impact on GPP predictions within different AutoML frameworks (Tramontana et al., 2016; Joiner and Yoshida, 2020). Each set consisted of different features that could explain the variation in GPP. The minimal set of remotely sensed variables (RS minimal) included surface reflectance from seven MODIS visible to infrared bands and PAR, which largely reflect the ability of the vegetation canopy to intercept solar radiation for photosynthesis. The RS set included all remotely sensed variables and their products. Notably, compared to the RS minimal set, the RS set also included land surface temperature, evapotranspiration, and soil moisture, which provide an additional link to vegetation heat and water stress (Green et al., 2022; Stocker et al., 2018). Finally, the RS meteo set included all remotely sensed variables and, in addition, meteorological variables from the ERA5-Land reanalysis (see Table 2). Additionally, we replaced the MODIS reflectance bands, LAI, FPAR, and land cover products with their native 500 m resolution data in the RS set to evaluate the impact of satellite data spatial resolution on GPP estimation.

The explanatory variable sets can provide information about the importance of the input features for the performance of the upscaling frameworks. They are particularly important as many of the AutoML frameworks lack feature engineering algorithms and cannot select relevant features themselves.

2.3.2 Framework assessment

We used 5-fold cross-validation to train and evaluate the AutoML frameworks. Grouping the data by site helped us increase the independence between the folds and evaluate the models' ability to generalize spatially. Thus, a time series at one site could be assigned to only one fold and not split into training and test sets. In addition, stratification by land cover helped to distribute the folds similarly. We repeated the cross-validation 30 times with different random splits to evaluate the impact of partitioning the data on the final performance in our evaluation.

With H2O AutoML, Auto-sklearn, and AutoGluon, we selected popular frameworks for supervised regression problems on tabular data that support parallelization and a Python interface. Since AutoML is intended to work as an out-of-the-box solution, we kept the frameworks' configurations at default or recommended parameter values where it was possible and reasonable to do so. Moreover, we set each framework to optimize for the root mean squared error (RMSE) and limited the resource usage during training to 600 CPU minutes per CV fold (30 min on 20 CPUs) and 64 GB of memory.

We used the RMSE and the coefficient of determination (r^2) to evaluate the frameworks' performance by comparing the out-of-fold predictions to the ground truth values of GPP (Eq. A1). The latter aligns with the Nash–Sutcliffe model efficiency (Nash and Sutcliffe, 1970) used in some literature as a performance metric for the GPP prediction (e.g., Tramontana et al., 2016). In addition to obtaining performance metrics for the total time series prediction, we decomposed the time series to evaluate the performance in different spatial and temporal domains. We computed the components as follows: we obtained trends by linear regression of the entire time series (using the slope for evaluation with RMSE and r^2), seasonality (mean seasonal cycle) by monthwise averaging, and anomalies as their residuals after detrending and removing seasonality. Furthermore, we calculated an across-site variability from the multi-year mean at each site. For this analysis, we considered only sites with a minimum of

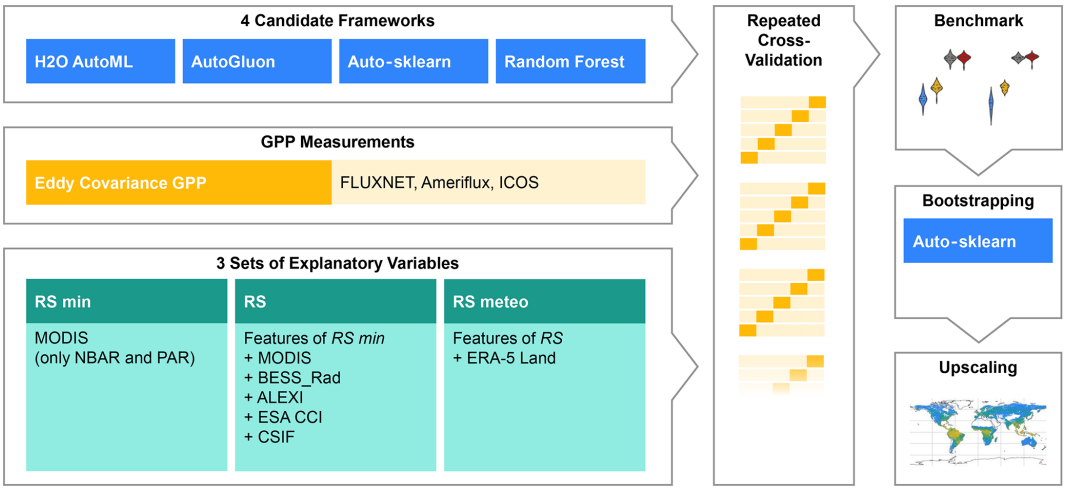

Figure 3. Experiment setup. We trained and evaluated Auto-sklearn, H2O AutoML, AutoGluon, and random forest together with three sets of explanatory variables in repeated cross-validation on GPP data from eddy covariance measurements. Then, we trained Auto-sklearn in a bootstrap aggregation to produce global wall-to-wall GPP maps. The abbreviations of the explanatory variable sets translate as follows: RS for remotely sensed data and meteo for meteorological data.

Table 2. Explanatory variable sets and associated datasets.

Explanatory variable	RS minimal	RS	RS meteo
Reflectance (nadir-BRDF adjusted; NBAR), bands 1–7	•	•	•
PAR	•	•	•
Diffuse PAR		•	•
RSDN		•	•
FPAR		•	•
LAI		•	•
Land surface temperature (day)		•	•
Land surface temperature (night)		•	•
ET		•	•
Soil moisture		•	•
SIF		•	•
Instantaneous SIF		•	•
Land cover (biome)		•	•
Total precipitation			•
Total precipitation (3-month lag)			•
Temperature			•
Vapor pressure deficit (VPD)			•

24 months of measurements to minimize the error from sites with just a few measurements, leaving us with 211 sites. When calculating trend metrics, we only considered sites with at least 60 months of measurements for our trend evaluations. Time series anomalies were detrended only when this minimum was reached; otherwise, we simply removed the seasonal component from the time series.

Moreover, we tested how the average ranked performance of each framework compared to the other frameworks. We calculated the performance ranks within each repeated cross-validation and obtained an average rank for each framework. Using the Friedman test, we tested for statistically significant differences in the rank distribution, evaluating the null hypothesis of no significant differences with a significance level of 0.01. We then used the Nemenyi post hoc test to find frameworks with significant differences in mean rank while adjusting for type I error inflation by using a family-wise error correction. We rejected the null hypothesis (no significant difference between the two frameworks) if the difference between the average ranks exceeded a critical difference (CD), which depends on the critical value of the Studentized range distribution (Demšar, 2006).

2.3.3 GPP upscaling

We used Auto-sklearn with the RS explanatory variable set to upscale the eddy covariance measurements to a global scale, as this combination of framework and explanatory variables performed best in the benchmark. We trained 30 models in a bootstrap aggregation approach, where each bootstrap was sampled with replacement to a size of 80 % of the total number of sites. We kept the time series grouped by site but removed the land cover stratification. This technique allowed us to estimate GPP as the mean of the bootstrapped predictions and provided a sampling error (standard error of the mean) as a spatially distributed uncertainty estimate for the model prediction. We produced global GPP and standard error maps at a resolution of 0.05° in monthly frequency from 2001 to 2020, which we compared with the two ML-based reference datasets FluxCom v6 (RS only, based on data from MODIS collection 6) (Jung et al., 2020) and FluxSat (Joiner and Yoshida, 2020).

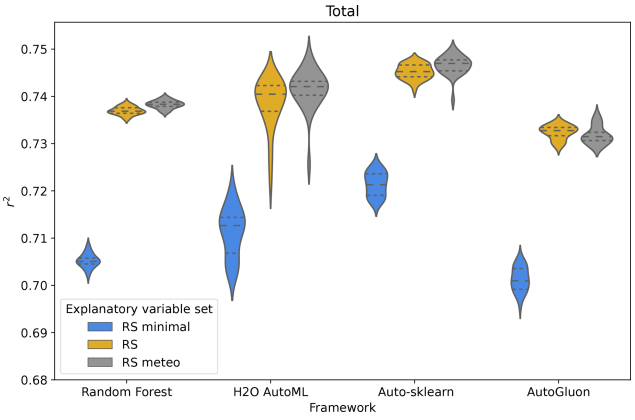

Figure 4. Overall framework performance, expressed as the coefficient of determination (r^2), for the candidate frameworks and the three different explanatory variable sets. Each distribution belongs to one framework and one set of explanatory variables and results from the repeated cross-validations, for each of which one r^2 value is calculated over the predictions at all sites.

3 Results

3.1 AutoML framework performance

In general, we found that all frameworks perform in a close range of coefficients of determination (r^2), explaining on average between 70 % and 75 % of the variation in eddy covariance GPP measurements. However, the performance depends on the framework used and the selection of variables. Examining the distribution of r^2 values for the different repeated cross-validations, we can see that Auto-sklearn performs best, followed by H2O AutoML, random forest, and AutoGluon, in predicting monthly GPP (Fig. 4). Auto-sklearn achieved the highest r^2 among the four frameworks for all explanatory feature sets. A similar pattern is observed for trends, seasonality, across-site variability, and anomalies (Fig. 5). Note that we removed one outlier for H2O AutoML trained on the RS variable set, which deviated more than 5 standard deviations from the mean value due to very low performance in one CV fold.

Auto-sklearn's superior performance is primarily due to its ability to capture seasonal components and across-site variability (Fig. 5). When trained on RS explanatory variables, Auto-sklearn achieved average r^2 values of 0.7452 ± 0.0003 overall and 0.483 ± 0.002 for trends, 0.8142 ± 0.0003 for seasonalities, and 0.689 ± 0.001 for across-site variability. However, all models struggle to reproduce the monthly anomalies, explaining less than 11 % of the variability (Auto-sklearn: 10.40 ± 0.04 %). Uncertainties are reported as the standard error of the mean of all cross-validation results.

Using the Friedman test, we found that the four ML frameworks are statistically different in their performance in predicting monthly GPP as well as its trends, seasonality, anomaly, and across-site variability (p value < 0.01). However, their difference in performance is marginal. The Nemenyi post hoc test shows that for the RS explanatory variables, Auto-sklearn achieves the highest average rank with statistical significance among all frameworks for monthly GPP and all its components (Fig. 6a). For the prediction of anomalies, we could not find a significant difference in the average rank between Auto-sklearn and H2O AutoML. Trends were predicted by all AutoML frameworks without significant differences in rank. Random forest and AutoGluon perform the worst, while they are not statistically different in predicting across-site variability and seasonalities.

The selection of explanatory variables had a significant impact on the performance of the frameworks. Models with only surface reflectance and PAR (RS minimal) explained the least amount of GPP variability (70 %–72 %) (Fig. 4). The greatest improvement occurred with the RS set when information on SIF, FPAR, LAI, LST, ET, soil moisture, and biome type was included. The RS set increased r^2 on RS minimal by about 0.02 for all frameworks, with sizable improvements in predicting trends and anomalies (Fig. 5). Meteorological variables except for AutoGluon slightly improved the prediction of monthly GPP by better explaining spatial variability, trends, and anomalies (Fig. 5). However, statistical tests of model ranks showed no significant advantage in the rank of the RS meteo over the RS set of explanatory variables in any of the decomposed time series features and frameworks (Fig. 6b). The RS set outperformed RS minimal in predicting GPP and all of its spatiotemporal components. Except for the performance of random forest on across-site variability, trend, and anomalies, RS was always the best-performing variable set or insignificantly different from the best-performing variable set. In addition, we evaluated whether vegetation indices (VIs) could improve the

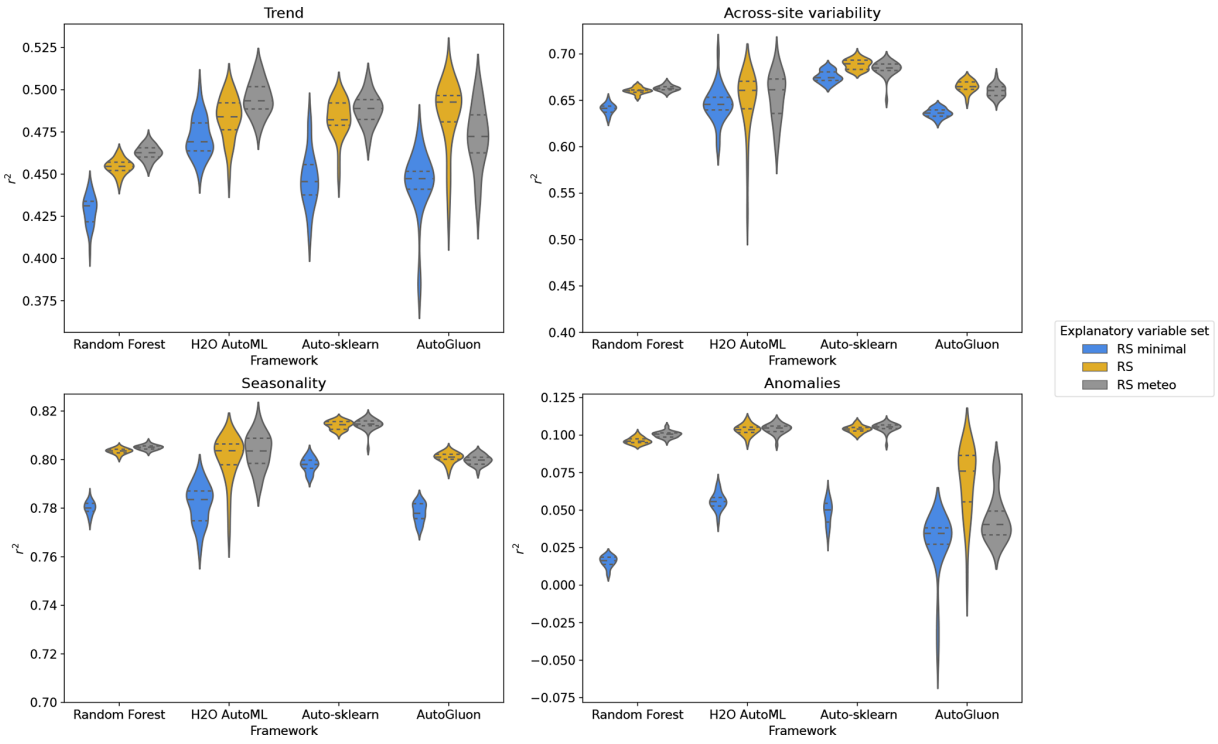

Figure 5. Evaluation of the temporally and spatially decomposed time series expressed as the coefficient of determination (r^2). Each distribution belongs to one framework and one set of explanatory variables and results from the repeated cross-validations, for each of which one r^2 value is calculated over the predictions at all sites. The r^2 values for seasonality and anomalies were calculated from seasonal cycles and anomalies at monthly granularity, while those for trend and across-site variability were calculated from one trend and mean value per site, respectively.

performance of the variable sets, but no improvements were found beyond the RS minimal dataset (Table A1).

To determine which explanatory variable was most effective for predicting GPP, we evaluated the permutation importance of the variables for the Auto-sklearn framework. Permutation importance is the decrease in prediction performance on the test dataset when one of the variables is randomly shuffled to break its relationship with the target variable. To deal with collinearity among the explanatory variables (Fig. A1), we first clustered them based on their average mutual Pearson correlation coefficient, regardless of their data source or ecological function. Variables with an average correlation greater than 0.7 were clustered and permuted together, resulting in clusters focused around specific meteorological characteristics (e.g., precipitation, temperature), vegetation properties, or combinations of reflectance bands but also combining features that are not directly biophysically related (Figs. A2 and A3).

Our results show the largest decrease in r^2 of Auto-sklearn RS when removing the cluster of SIF, LAI, and FPAR, followed by PAR, RSDN, LST, and ET (Fig. 7). The other variables do not substantially reduce the framework performance. Trained on RS meteo, Auto-sklearn's variable importance gives a similar picture despite slightly different clusters due to the inclusion of the meteorological variables. Again, the cluster of SIF, LAI, and FPAR shows by far the highest importance, followed by the PAR, RSDN, ET, and temperature-related variables (Fig. 7). The meteorological variables temperature, VPD, and precipitation are generally in clusters of lower importance, as are the MODIS NBAR features. In contrast, the RS minimal product shows the highest variable importance for the visible NBAR spectrum, followed by NIR and PAR in descending order. The shortwave infrared (SWIR) bands are hardly used in any setup.

Furthermore, we grouped the predictions by site and evaluated the site-level r^2 for each land cover type for Auto-sklearn with RS explanatory variables (Fig. 8). EBF and SH sites show low r^2 (median r^2 – 0.38 and 0.33, respectively) with substantially higher variance, whereas MF and DBF could be predicted with high quality (median r^2 0.84 and 0.87, respectively). Regarding anomaly estimation, EBF and WET show significantly lower r^2 values (median r^2 0.04 and 0.01, respectively). Furthermore, our analysis indicated that models tended to exhibit a significant positive bias when predicting small GPP values (in the lowest quartile), while displaying a negative bias for large GPP values. This implies an overestimation of small GPP and an underestimation of large GPP values by the models.

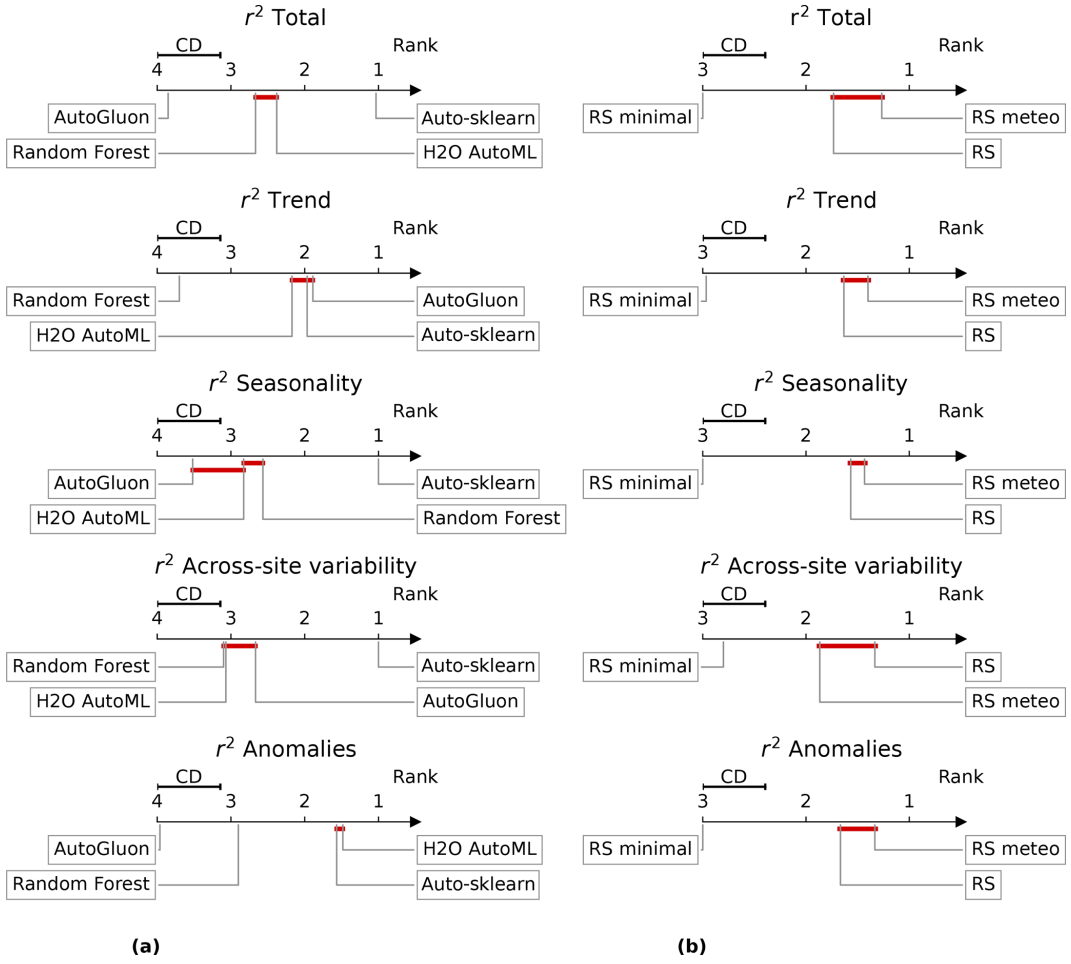

Figure 6. Critical difference (CD) diagrams (Demšar, 2006) for the ranks of the frameworks and variable sets, which are typically used to compare the performance of multiple algorithms on multiple problems (in this case, repeated cross-validations). The graphs rank the performance of different framework–variable combinations on the x axis, with one being the best rank. The ranks shown are the average ranks from all repeated cross-validations for each of the frameworks and variable sets. The performance (r^2) is given for predicting total GPP and its different spatial and temporal components: trend, seasonality, anomalies, and across-site variability. We evaluated whether the ranks are statistically significantly different from each other using the critical difference (CD) obtained from a Nemenyi post hoc test. If the difference between the ranks is less than the CD, we assume a nonsignificant difference in ranks, indicated by the red crossbar between the rank markers. On the left side **(a)**, the ranks of the frameworks trained on the RS explanatory variables are shown. On the right side **(b)**, the ranks of Auto-sklearn trained on different sets of explanatory variables are shown.

Finally, we examined the effect of including higher-resolution data in the explanatory data. Replacing the MODIS reflectance bands, LAI, FPAR, and land cover products with their 500 m resolution counterparts resulted in significant improvements in r^2. We tested this behavior for Auto-sklearn with the RS variable set. The prediction r^2 was significantly higher than for the lower-resolution data product in all aspects except trend with 0.8164 ± 0.0005 overall and 0.444 ± 0.003, 0.787 ± 0.002, 0.8723 ± 0.0005, and 0.3094 ± 0.0006 for trend, across-site variability, seasonality, and anomalies, respectively (Fig. 9).

3.2 Analysis of Auto-sklearn pipelines

We investigated the different components (base models and preprocessing algorithms) of the Auto-sklearn framework, which was trained on the RS variable set in the repeated cross-validation (see Fig. A4 for the model run statistic). For every fold in each of the repeated cross-validations, we considered the best-performing model of each base model type and min–max scaled their RMSE to a scale from 0 to 1. The scaling accounts for the different predictability of the test data in the respective fold. We then took the mean across all folds within each repetition of the cross-validation and each base model type, resulting in a distribution of scaled RM-

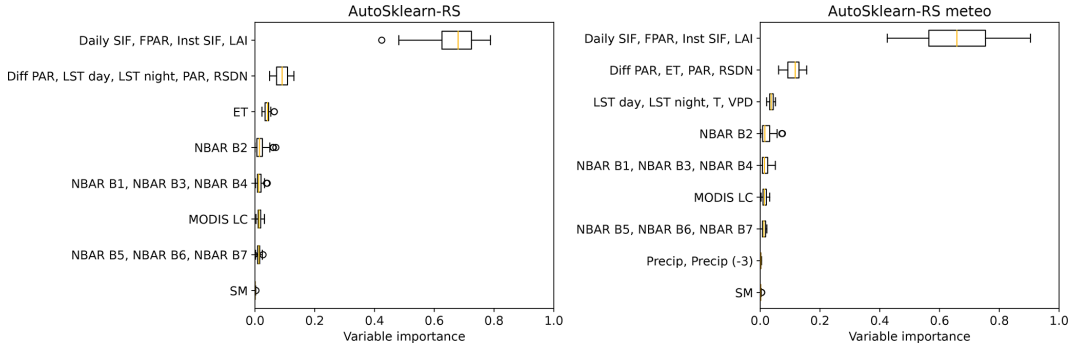

Figure 7. Permutation importance for different explanatory variables with the Auto-sklearn framework and RS and RS meteo variable sets. The variables are grouped into clusters of colinear variables regardless of data source or ecological function. The importance is the decrease in r^2 at test time when the variables of the corresponding cluster are randomly shuffled. The variables include the MODIS NBAR bands (red, NIR, blue, green, and three SWIR bands), land surface temperature (LST), leaf area index (LAI), photosynthetically active radiation (PAR), fraction of absorbed PAR (FPAR), diffuse PAR (Diff PAR), daily and instantaneous solar-induced fluorescence (SIF), surface downwelling shortwave flux (RSDN), soil moisture (SM), evapotranspiration (ET), precipitation (Precip), temperature at 2 m height (T), vapor pressure deficit (VPD), and precipitation with a 3-month lag (Precip (-3)). The distribution results from the repeated cross-validations, for each of which one r^2 value is calculated over the predictions at all sites.

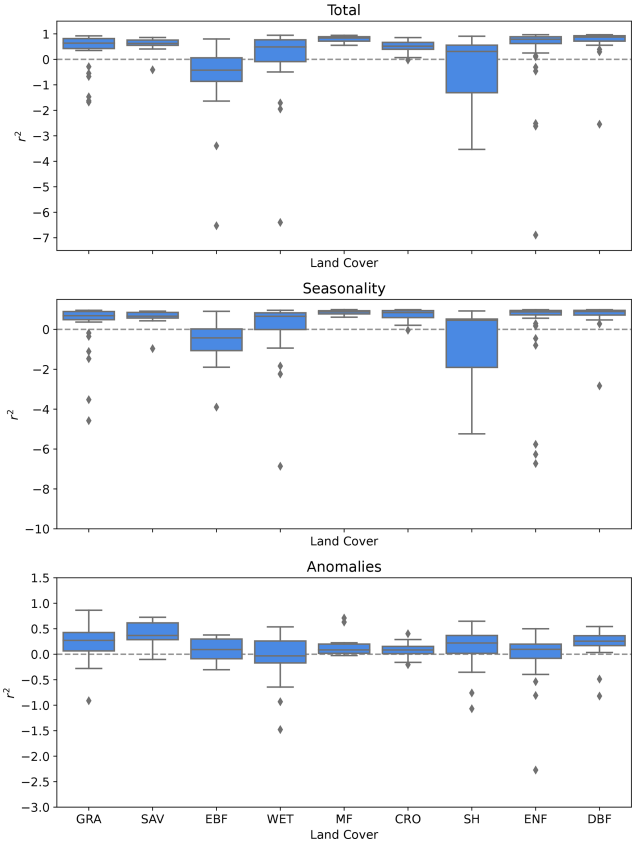

Figure 8. Distribution of r^2 values for the GPP prediction by Auto-sklearn with RS explanatory variables for different land cover types. Shown are the overall performance and performances for seasonality and anomalies.

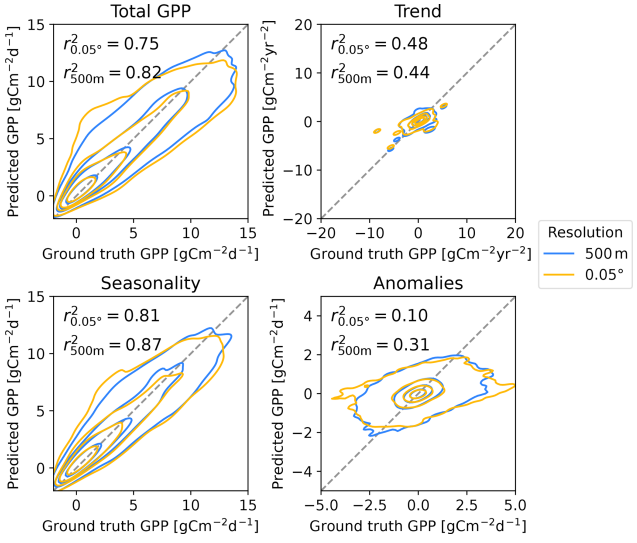

Figure 9. Comparison of the predicted 0.05° product and the one with 500 m resolution from Auto-sklearn ensemble averages and the RS variable set. The latter shows higher r^2 values compared to the ground truth GPP estimates from FLUXNET, AmeriFlux OneFlux, and ICOS. We refer to GPP measurements derived from eddy covariance at the flux tower locations as ground truth.

SEs for each base model type (Fig. 10). We also considered whether these models preprocessed the training data or not.

The base models achieving the lowest scaled RMSE were ensembles of weak learners, such as extra trees, random forest, gradient boosting, or AdaBoost. These models could, by themselves, achieve the best predictions of GPP. That, however, does not suggest that they were necessarily used in the final model ensemble constructed by Auto-sklearn. The ensemble selection algorithm (forward stepwise model selec-

tion) in Auto-sklearn, which creates the model ensembles, recursively adds the base models that improve the RMSE of the ensemble prediction most in combination with the models already part of the ensemble (Caruana et al., 2004). Hence, a model showing a low RMSE by itself does not need to be beneficial to the ensemble of models ultimately used by Auto-sklearn.

3.3 Global GPP maps

From the bootstrap aggregation of the Auto-sklearn framework with RS features, we predicted global GPP with wall-to-wall coverage, resulting in 30 predictions for the entire period from 2001 to 2020 in monthly intervals. In addition, we applied land–sea and vegetation masks to the prediction, similar to previous research (Tramontana et al., 2016; Joiner et al., 2018).

Mean GPP for 2001–2020 (Fig. 11) showed high values for tropical climates in low latitudes, such as the Amazon region, Southeast Asia, and central Africa, with maximum GPP values for the EBF land cover. Conversely, low GPP appears in high latitudes and SH, SAV, and GRA regions.

Again, we decomposed the local time series into trends, seasonality, and anomalies (Fig. 11). The amplitude of the seasonal component exhibits significant regional differences. Mid-latitude regions in the Northern Hemisphere show high amplitudes, covering the central and eastern US, Europe, parts of Russia, and northeastern China. In contrast, low-latitude regions have low GPP amplitudes. The data show significant trends ($p < 0.05$) over the observation period with positive clusters, especially for eastern China and western India, while negative trends are less pronounced. The bootstrapped Auto-sklearn framework shows clusters of high GPP anomalies in, e.g., parts of South America (especially eastern Brazil and Argentina), east Africa, and southeast Australia. Land cover in these areas does not follow a consistent pattern but is often dominated by CRO, SH, and GRA.

In addition to the GPP prediction, we produced a sampling error estimate by calculating the average standard error across all bootstraps for each location and time (Fig. 12). We observed high relative errors in low-GPP regions, high-latitude regions (e.g., with temporary snow cover), and arid SH regions. The distribution of standard errors relative to the bootstrap mean peaks near 0 and ends in a long tail towards higher values for all biomes (Fig. 13a). However, the distribution of sampling uncertainty in GPP varies among land cover classes, ranging from low medians for EBF (0.5 %) and SAV (0.8 %) up to higher medians for ENF (4.0 %) and SH (6.9 %).

3.4 Comparison to reference data

We compared the upscaled results of total GPP from our Auto-sklearn RS prediction with GPP datasets FluxCom v6 (Tramontana et al., 2016) and FluxSat (Joiner et al., 2018) at 10 000 random sample locations. When tested with a Mann–Whitney U test, our predictions show significantly higher agreement (p virtually 0) with FluxSat than with FluxCom (Fig. 13b). In our prediction, 51 % of the samples explain more than 80 % of the variation in FluxSat, while this is the case for only 17 % of the samples in FluxCom. Thus, Auto-sklearn shows good agreement with the GPP patterns predicted by FluxSat, whereas it deviates more strongly from the FluxCom product.

4 Discussion

4.1 AutoML framework performance

The results demonstrate the closeness of the overall predictive performance of the evaluated frameworks and the baseline random forest. Despite the different complexity of the model architectures, the frameworks capture a similar fraction of the variability in the GPP measurements. Framework choice does not appear to be a major factor in this experimental setup, resulting in only a low difference in r^2. These findings align with previous research on applying classical ML models (Tramontana et al., 2016).

The performance differences between the frameworks are statistically significant but slight. Auto-sklearn consistently outperforms H2O AutoML, AutoGluon, and random forest. The framework is based on ensemble prediction, which can exploit the different advantages of each base model. The evaluation of base models used by Auto-sklearn outlines the applicability of various ML model types for predicting GPP. It is evident that ensembles of weak learners, such as extra trees or random forest, are generally favorable for this task. These models can be promising for GPP prediction either in a standalone implementation or as part of a model ensemble. The performance comparable to H2O AutoML and AutoGluon shows furthermore that implementing feed-forward neural networks does not necessarily lead to performance improvements. Low performance of AutoGluon, even when compared to random forest, may relate to the lack of hyperparameter tuning. However, the differences between frameworks are challenging to explain, as the reasons for the frameworks' results are obscured by their black box character.

Auto-sklearn trained on RS explanatory variables tended to overestimate small GPP values, while underestimating large GPP values. This behavior was already observed in the FluxCom (RS), FluxSat, and several light use efficiency models (Yuan et al., 2014; Joiner et al., 2018). It has also been shown for the early MODIS GPP product (Running et al., 2004), where the overestimation was attributed to an artificially high FPAR, while the underestimation was related to low-light-use efficiency in the MODIS algorithm (Turner et al., 2006). Another reason could be the strong reliance of the Auto-sklearn framework on tree-based models (Fig. 10).

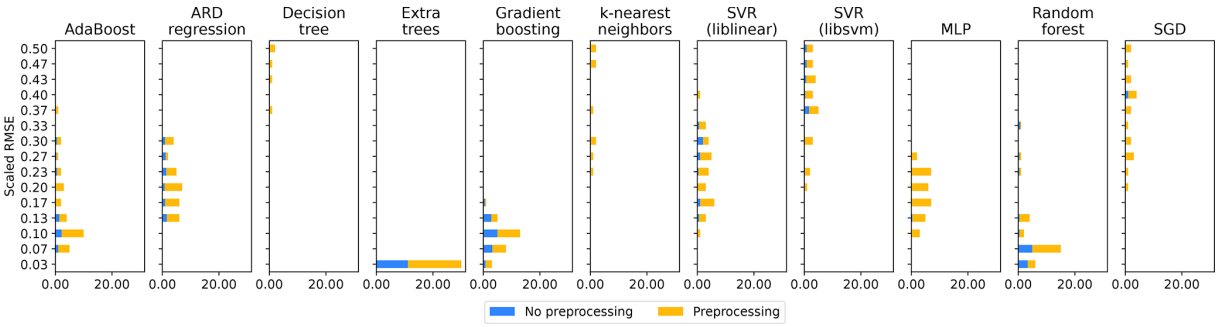

Figure 10. Performance of Auto-sklearn base models and feature preprocessors. The chart shows the distribution of the mean RMSE for each base model type across all folds within each repetition of the cross-validation. We considered only the best-performing models for each model class within each fold. The RMSE is min–max scaled from 0 to 1 within each cross-validation fold to account for variations in the data's predictability depending on the data's split. The use of preprocessing algorithms is shown as colors in the proportions of their usage in each bin (detailed preprocessing methods in Fig. A5).

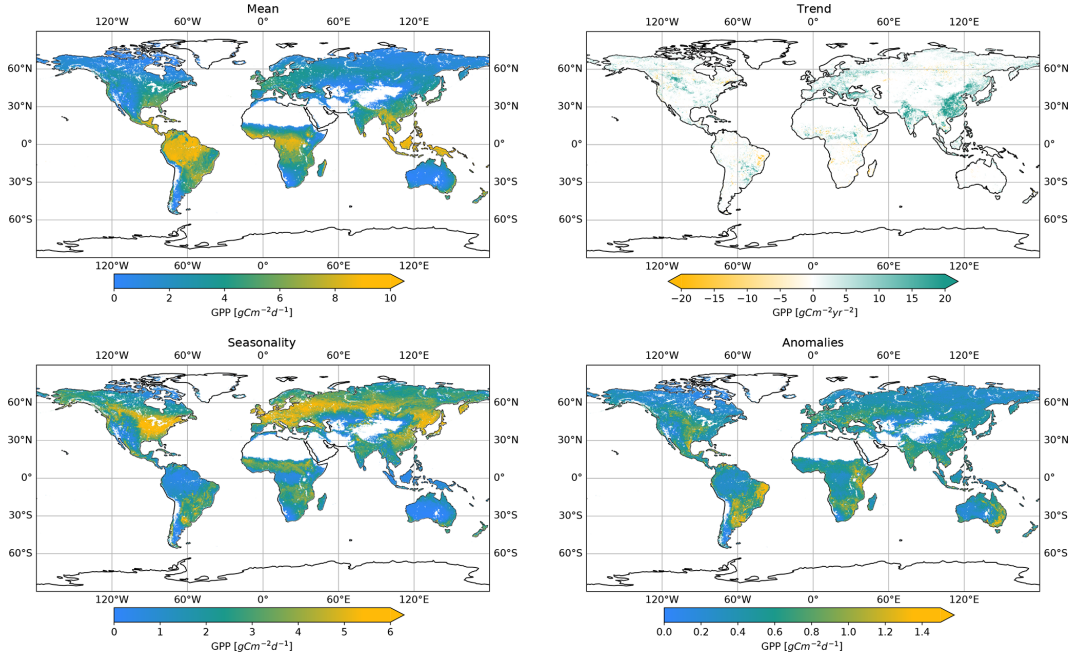

Figure 11. Total GPP, amplitude of seasonality, trend, and anomalies of prediction with Auto-sklearn trained on remotely sensed data (RS dataset) in a bootstrap aggregation of 30 bootstraps. The mean was calculated at each location over all bootstrapped predictions and the entire time series. The seasonality is displayed as the amplitude of the monthwise average. Trends were calculated as the slope from an ordinary least squares linear regression over time and masked so that only significant trends were included ($p < 0.05$). The anomalies are shown as the standard deviation of the residuals after subtracting the seasonal and trend components from the time series.

These models are constructed by recursively partitioning the feature space into small regions into which they fit a simple model, which limits them in their ability to extrapolate beyond the range of target values already observed. Furthermore, our predictions showed differing prediction quality at the land cover level, which might result from biome-specific circumstances and the availability of measurement sites. For example, biomes with a pronounced seasonal cycle, such as DBF or MF, exhibit high overall r^2, whereas EBF and WET show large variability that the model could not capture. In addition, variability within a land cover type could affect the performance assessment, such as for SH, which includes both arid and subarctic shrublands.

Finally, it is crucial to note that the r^2 metric only expresses how well a framework can reproduce measurements from the measurement samples, which are limited in underrepresented areas. We grouped data by site and applied a land cover stratification during the CV to increase independence between the folds. That, however, does not prevent sites from being repeatedly selected for validation during the

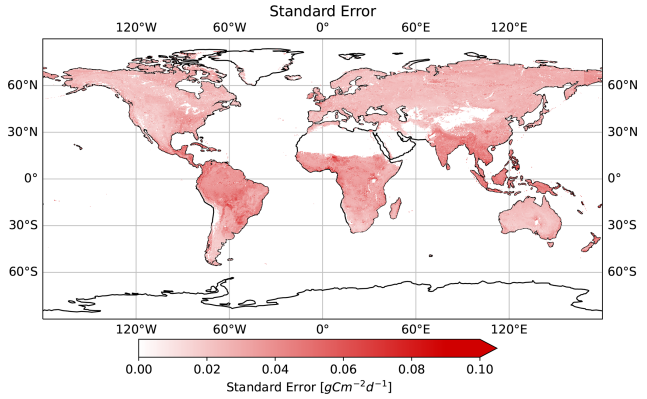

Figure 12. Absolute standard errors from the bootstrap aggregation. For relative values, see Fig. A6.

repeated CV, which can inflate the performance metric and reduce variance. It also cannot account for spatial autocorrelation. This affects the assumption of independence and identical distribution for train and test folds, which is crucial for obtaining realistic CV results. Violating these requirements can lead to overestimating model performance and inflating map accuracies, yet it is commonly done in data upscaling efforts (Roberts et al., 2017; Ploton et al., 2020). More training data with better geographic representation could help mitigate these shortcomings and could lead to more robust predictions, model evaluations, and potentially higher model performance.

4.2 Importance of explanatory variables

AutoML is a powerful approach for assessing the importance of the variables on model performance since it selects the optimal base models and constructs optimal pipelines independently for each feature set under consideration. This means that no subjectivity bias is introduced into assessing variable importance, e.g., by pre-selecting specific algorithms that are expected to perform well on a particular task or set of explanatory variables. This could increase the quality of the reported importance, especially as features in GPP prediction often exhibit severe intercorrelations. Importantly, variable importance is model-specific, meaning it can indicate which variable is most effectively used by a particular model, but it does not directly indicate the intrinsic predictive value of a variable. Furthermore, it may depend on the choice of temporal and spatial scales and data quality, given that many of the input features are themselves model outputs.

The frameworks' performance depends significantly on the choice of predictive features on which they are trained. The results show that while the seven NBAR bands and PAR from the RS minimal variable set provide the model with sufficient information for a GPP prediction, the full set of RS variables adds additional information that all the frameworks can exploit. The additional variables in the RS variable set, such as SIF, LAI, FPAR, ET, LST, SM, and plant function type, appear to include important environmental forcings and structural variables that provide a marginal advantage over the variables on only vegetation structure and radiation in RS minimal (Green et al., 2019; Stocker et al., 2019; Xu et al., 2020). For example, environmental stress, such as heat waves and droughts, often causes instantaneous reductions in GPP. However, the response of vegetation greenness to these stressors is typically slower and may only become apparent if the stress persists for a sufficient duration (Orth et al., 2020; Zhang et al., 2016; Smith et al., 2018; Yan et al., 2019). In such cases, relying solely on surface reflectance may not sufficiently capture the variability of GPP.

Including the meteorological explanatory features (ERA5-Land) in the training data does not significantly improve the prediction quality for any of the frameworks. This implies that meteorological data may not contain additional information that the machine learning frameworks in this study can effectively use to predict GPP. A possible explanation could be the mismatch between reanalysis and site meteorology. The coarse resolution and large uncertainties in the reanalysis data may result in a poor representation of the flux tower footprints, which are often smaller than one pixel of the reanalysis data, leading to uncertainties in the modeling. For example, Joiner and Yoshida (2020) showed that using site-measured meteorological data instead of reanalyzed data significantly improved the performance of GPP predictions. At the monthly scale, the RS variable set may already encode information about the instantaneous environmental stress from adverse meteorological conditions through, for example, LST, ET, and soil moisture, which are important controls on GPP (Bloomfield et al., 2023). Further studies could potentially assess these uncertainties by comparing models trained with tower meteorological data to gridded reanalysis datasets.

The permutation importance of explanatory variables provides further insight into which variables Auto-sklearn uses and which are indifferent to the framework. Our results show that both RS- and RS-meteo-trained Auto-sklearn frameworks rely primarily on features of canopy structure (LAI, FPAR), proxies for photosynthetic activity (SIF), and ET, which strongly couples with GPP in favorable environmental conditions. Meteorological information, such as temperature and VPD, is less relevant for the model prediction. This suggests that the insignificant changes in performance between RS and RS meteo may be related to a small additional contribution of meteorological conditions to the prediction of monthly GPP beyond what is already provided by vegetation structure and PAR. Soil moisture was also found to have minimal influence overall, which might be partly due to uncertainties and noises in the remote sensing soil moisture data and due to its coarse spatial resolution. It is also important to note that previous studies have demonstrated the importance of soil moisture from SMAP in predicting GPP in water-limited ecosystems (Dannenberg et al., 2023; Kannen-

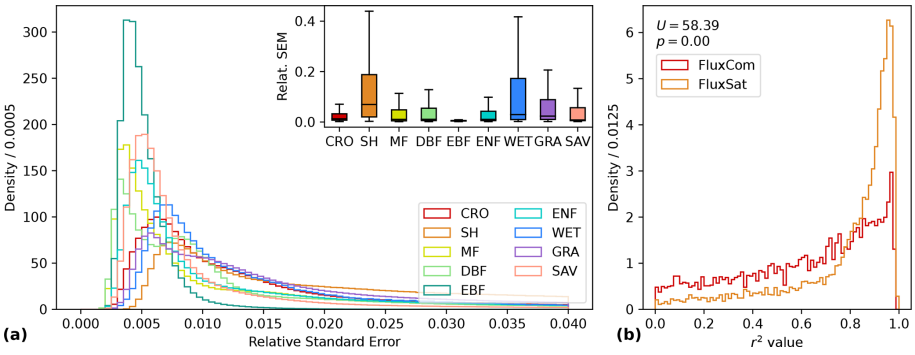

Figure 13. Histogram of the relative standard error of the mean (SEM) by land cover class during the entire observation period **(a)** and distribution of r^2 values for total GPP of the upscaled GPP Auto-sklearn product with RS variables compared to the FluxCom v6 and FluxSat datasets **(b)**. For the latter, GPP is sampled at 10 000 random locations and compared in a Mann–Whitney U test.

berg et al., 2024). The performance difference between RS minimal (NBAR and PAR only) and RS variables seems to be driven at least partly by features that are themselves model outputs based on MODIS NBAR, i.e., SIF, LAI, and FPAR. We grouped the variables into clusters with high correlation to improve the interpretability of the importance measures. However, we could not completely eliminate correlations between clusters. High correlations between, for example, PAR and LST and ET and PAR, as well as lower correlations between other variables, could not be taken into account and introduced further uncertainty in the reported variable importance.

The ability of the frameworks to reproduce GPP patterns and the corresponding variable importance must be evaluated in light of the choice of temporal resolution. In this study, we evaluated machine learning upscaling of monthly GPP dynamics, which are dominated by light availabilities and seasonal changes in vegetation structures. However, at shorter timescales, such as hourly or daily, GPP is more closely aligned with diurnal and short-term variations in meteorological conditions such as temperature and VPD. Thus, these variables are likely more influential in predicting GPP at these higher frequencies (Frank et al., 2015; von Buttlar et al., 2018). Additionally, complex machine learning models may also offer greater benefits at harnessing the large data quantities involved in predicting GPP at hourly or daily scales. Further research is needed to benchmark machine learning algorithms and assess choices of environmental data in predicting GPP across different timescales.

We found that besides selecting an appropriate set of explanatory variables, the resolution of the data highly affects prediction outcomes. Including 500 m resolution data should reduce the mixed-pixel problem and match the flux towers' footprints better with the pixel size of the gridded datasets. This led to improvements in all time series components, with exceptional increases in r^2 for the estimation of anomalies. These results underscore the importance of spatial resolution and suggest the use of data with a resolution that better represents smaller landscape features and flux tower footprints, in contrast to our initial choice of 0.05° resolution in this study (Xiao et al., 2008; Yu et al., 2018; Chu et al., 2021).

4.3 Spatiotemporal patterns

The globally upscaled measurements could capture the variation in GPP in the ML-based FluxCom and FluxSat reference datasets reasonably well and resemble their total GPP patterns and seasonality (Tramontana et al., 2016; Joiner and Yoshida, 2021). However, the prediction could explain a significantly larger fraction of the variation in FluxSat than in FluxCom. Both datasets are based on MODIS-derived products, but the training sites we used show higher similarities to FluxSat than to FluxCom.

We observed several clusters of positive trends consistent with previous results and local studies (Chen et al., 2019; Wang et al., 2020; Schucknecht et al., 2013; Carvalho et al., 2020). However, the magnitude was lower than the reference dataset FluxSat (Joiner and Yoshida, 2021) and showed less frequent significant negative trends than predicted by FluxCom (Tramontana et al., 2016). The areas with high predicted GPP overlap with the highly productive regions in the tropics and mainly cover the EBF regions (Ahlström et al., 2015). In addition, we observed high seasonality, especially in CRO-dominated regions, which may be due to high productivity in maize, wheat, rice, and soybean cultivation and a profound seasonality, with a period of very low GPP after harvest. (Kalfas et al., 2011; Gray et al., 2014; Sun et al., 2021). High anomalies occurred in mainly temperate and semi-arid climates, the latter of which has also been shown to dominate the interannual variability of the global terrestrial carbon sink (Ahlström et al., 2015). Besides random variations included in the anomalies, reasons could be non-seasonal events, such as weather extremes or human interventions, coupled with a high turnover rate in dry vegetation. The patterns agree with FluxSat and exceed those that FluxCom models estimated.

4.4 Uncertainty

Predicting wall-to-wall maps from a non-representative distribution of measurement sites is challenging. A non-representative network of flux towers might fail to reproduce the main features of the underlying GPP population for the entire study area (Sulkava et al., 2011). Land cover types with less abundant eddy covariance measurements may potentially be estimated less reliably and could show a higher variation in GPP estimations. We used the standard error to estimate how robustly the frameworks react to different subsets (bootstraps) of data during the training process. Generally, high relative error values in low-GPP regions are expected due to the normalization of the error. However, SH, ENF, and regions adjacent to SNO and BAR also show an elevated error in absolute terms. The distributions (Fig. 13a) show similarities to the spread of r^2 values obtained from the framework benchmark (Fig. 8).

Higher standard errors may indicate that monthly remote sensing and modeled input data are better proxies for some ecosystems than others. For example, GPP can be predicted with low relative uncertainty for ecosystems with a high seasonal variation of biomass, such as croplands, broadleaf forests, and mixed forests. In contrast, predicting GPP in drylands can be more challenging. Drylands are highly sensitive to water availability, resulting in abrupt responses to precipitation and drought events (Barnes et al., 2021). They are characterized by high spatial heterogeneity and irregular temporal vegetation patterns, which are difficult to capture at our spatial and temporal resolution. Together with a low vegetation signal-to-noise ratio, these factors pose a considerable challenge for GPP remote sensing (Smith et al., 2019). In an attempt to assess the uniqueness of NEE measurements at FLUXNET sites, Haughton et al. (2018) showed that drier sites and shrubland sites had a higher discrepancy between locally and globally fit models and exhibited more idiosyncratic NEE patterns compared to others. Our results show a similar behavior, with higher model uncertainty for GPP in dryland and shrubland regions.

The results delineate that Auto-sklearn could not reliably infer a robust functional relationship in low-productivity regions, where it shows a significant positive bias. We suggest further research on ways to improve performance in low-GPP regions. One method that could potentially enhance the prediction is to include dummy measurement sites in the masked regions manually. These sites would constantly report zero GPP and could improve estimates in adjacent regions, such as arid zones or seasonally snow-covered areas, which are also less proportionately represented in the flux tower networks (Smith et al., 2019).

Finally, an additional limitation is introduced by the eddy covariance measurements themselves. We use nighttime-partitioned GPP, which is modeled as the difference between NEE and ecosystem respiration. While NEE and nighttime respiration are directly measurable, daytime respiration is modeled with a temperature response function, which extrapolates from nighttime respiration (Reichstein et al., 2005). Up to this point, it is not conclusively clarified how reliably this approach can be employed, considering that it is indifferent to some environmental stress factors and changes in respiration behavior between day- and nighttime (Wohlfahrt and Galvagno, 2017; Keenan et al., 2019; Tramontana et al., 2020). The inherent uncertainty and bias in the ground truth GPP data could be a potential cap to the performance we can obtain in our efforts to predict GPP.

5 Conclusions

We investigated whether and how automated machine learning (AutoML) frameworks can improve global upscaling of gross primary productivity (GPP) from in situ measurements using Auto-sklearn, H2O AutoML, AutoGluon, and a baseline random forest model in repeated cross-validation stratified by land cover. In addition, we evaluated different sets of explanatory variables for the GPP prediction from satellite imagery and ERA5-Land reanalysis data. Our results show that the AutoML frameworks can capture about 70 %–75 % of the monthly GPP variability at the measurement sites.

Auto-sklearn slightly but significantly outperformed the other frameworks across all sets of explanatory variables for total GPP, trends, seasonality, and anomalies. It did this by creating ensembles of base models and preprocessing algorithms that improved the prediction over individual machine learning models. The ensemble members were primarily models that combined weak learners, such as extra trees, AdaBoost, or random forests. However, the difference in performance was small compared to other frameworks and the random forest model, suggesting that the choice of framework may play only a minor role in improving GPP prediction performance.

We found that remotely sensed (RS) explanatory variables provided the best results in combination with the investigated frameworks. While only relying on the MODIS NBAR reflectance bands and PAR (RS minimal) provided the models with sufficient information for GPP prediction, considering other proxies of photosynthetic activity and canopy structure, such as solar-induced fluorescence, leaf area index, and fraction of absorbed photosynthetic activity, increased the performance of all models. Meteorological factors and soil water availability had less influence on the GPP prediction. Also, additional meteorological variables from ERA5-Land could not be used effectively by the models. In particular, the resolution of the satellite imagery played a significant role in prediction quality.

Finally, we used the best-performing framework (Auto-sklearn with RS explanatory variables) to upscale GPP to global wall-to-wall maps in a bootstrapping approach. The predictions are in good agreement with the FluxSat dataset and deviate significantly more from the FluxCom predic-

tions. The GPP product captures major spatial patterns for total GPP and trends but shows high uncertainty for low-GPP regions, where the predictions are positively biased. In general, prediction performance and sampling uncertainty are highly dependent on the land cover type.

In conclusion, AutoML can be a considerable technique for predicting and extrapolating GPP from in situ measurements. Automated creation of machine learning pipelines can facilitate the process of algorithm and feature selection, thereby avoiding biases in the modeling process. In addition, AutoML enables the exploration of a wide range of models and algorithms, uncovering potential relationships and patterns that may have been missed manually. However, we were unable to demonstrate that AutoML produces GPP predictions that are considerably more accurate and robust than classical ML models. In particular, the non-automated random forest model performed almost as well as Auto-sklearn. Researchers must carefully interpret and validate the results obtained through AutoML, ensuring that the models and features chosen are consistent with ecological knowledge and scientific understanding. Nevertheless, given the early stage of development, AutoML may be useful in the future to improve and accelerate research on GPP upscaling.

Appendix A

Equation (A1) is the coefficient of determination r^2, where y_i is the observed value, $\hat{y}_i$ the modeled value, and $\overline{y}$ the observed average over all N values.

$$r^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{A1}$$

Table A1. Overall framework performance. Shown are the mean r^2 values with the corresponding error of the mean, averaged over all cross-validation repetitions. Additionally to the three predictor variable sets, we added the following vegetation indices (VIs) to each variable set to evaluate if they improve the performance: the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), green chlorophyll index (GCI), normalized difference water index (NDWI), near-infrared reflectance of vegetation (NIRv), and kernel NDVI (kNDVI).

Variable set	Random forest	H2O AutoML	AutoSklearn	AutoGluon
RS minimal	0.7052 ± 0.0003	0.7112 ± 0.0009	0.7214 ± 0.0005	0.7013 ± 0.0005
RS minimal (incl. VI)	0.7193 ± 0.0002	0.7166 ± 0.0007	0.7261 ± 0.0004	0.7097 ± 0.0007
RS	0.7369 ± 0.0002	0.739 ± 0.001	0.7452 ± 0.0003	0.7324 ± 0.0003
RS (incl. VI)	0.7352 ± 0.0002	0.7383 ± 0.0004	0.7437 ± 0.0003	0.7315 ± 0.0002
RS meteo	0.7383 ± 0.0002	0.7416 ± 0.0008	0.7214 ± 0.0004	0.7318 ± 0.0004
RS meteo (incl. VI)	0.7356 ± 0.0002	0.7402 ± 0.0005	0.7201 ± 0.0003	0.7310 ± 0.0002

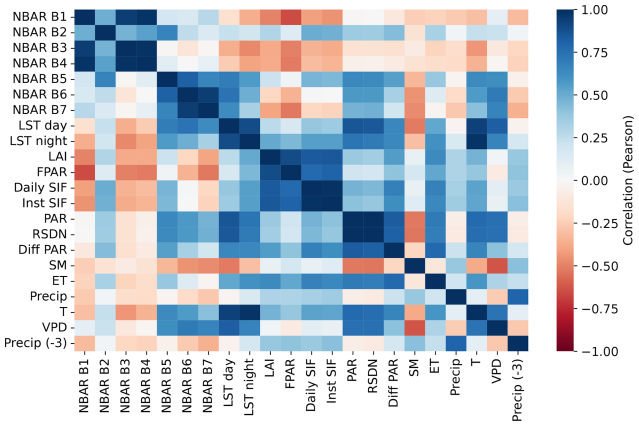

Figure A1. Pearson correlation matrix between the scalar explanatory variables, including the MODIS NBAR bands, land surface temperature (LST), leaf area index (LAI), photosynthetically active radiation (PAR), fraction of absorbed PAR (FPAR), diffuse PAR (Diff PAR), daily and instantaneous solar-induced fluorescence (SIF), surface downwelling shortwave flux (RSDN), soil moisture (SM), evapotranspiration (ET), precipitation (Precip), temperature at 2 m height (T), vapor pressure deficit (VPD), and precipitation with a 3-month lag (Precip (-3)).

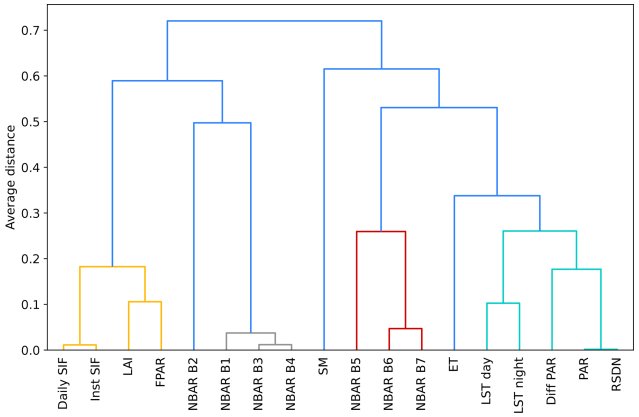

Figure A2. Dendrogram for clustering the explanatory variables of the RS set. The variables are clustered after their average distance, which is one minus the absolute of the Pearson correlation coefficient. See Fig. A1 for variable abbreviations.

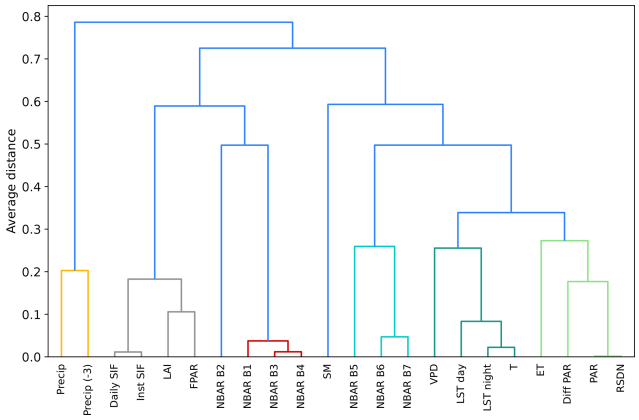

Figure A3. Dendrogram for clustering the explanatory variables of the RS meteo set. The variables are clustered according to their average distance, which is 1 minus the absolute of the Pearson correlation coefficient. See Fig. A1 for variable abbreviations.

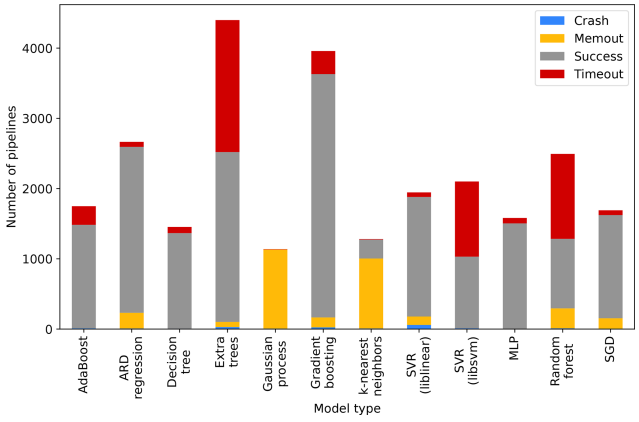

Figure A4. Run statistics of the Auto-sklearn base models. The four statuses show how many base models succeeded or failed during training due to insufficient memory, training time, or other unknown reasons. Only the successful models were used for the configuration of Auto-sklearn.

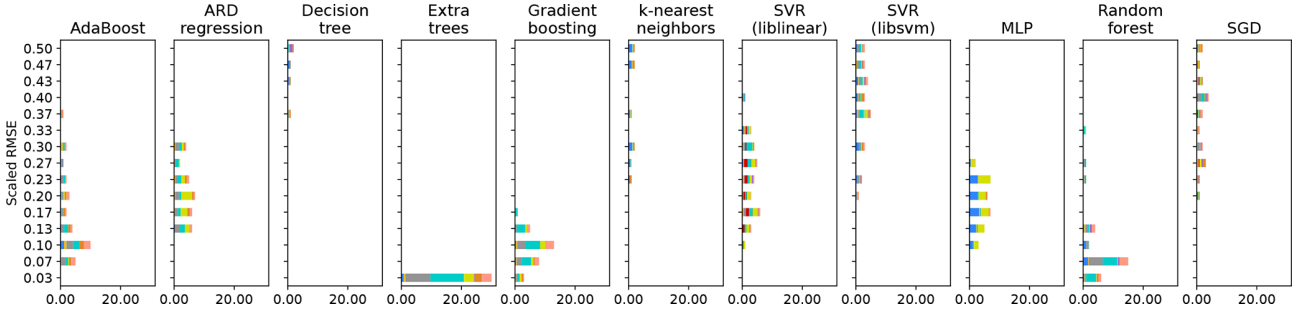

Figure A5. Detailed use of preprocessing algorithms by Auto-sklearn. The chart shows the distribution of the mean RMSE for each base model type across all folds within each repetition of the cross-validation. We considered only the best-performing models for each model class within each fold. The RMSE is min–max scaled from 0 to 1 within each cross-validation fold to account for variations in the data's predictability depending on the data's split. The use of preprocessing algorithms is shown as colors in the proportions of their usage in each bin.

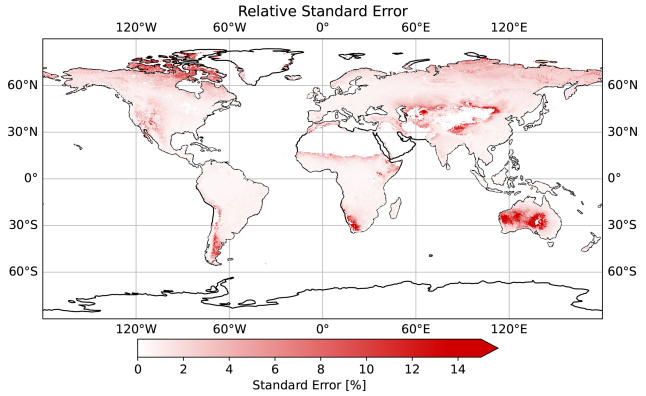

Figure A6. Relative average standard error, normalized by the mean GPP prediction.

Code availability. The code can be found at https://doi.org/10.5281/zenodo.8262618 (Gaber, 2023).

Data availability. The FLUXNET2015 data can be downloaded through the FLUXNET data portal (https://fluxnet.org/data/fluxnet2015-dataset, Pastorello et al., 2022). The AMERIFLUX data are available through the AMERIFLUX network (https://ameriflux.lbl.gov/data/flux-data-products, AmeriFlux Management Project, 2022). ICOS data can be accessed through the ICOS web portal (https://doi.org/10.18160/2G60-ZHAK; Warm Winter 2020 Team and ICOS Ecosystem Thematic Centre, 2022). All MODIS products are available through the NASA EOSDIS Land Processes DAAC (https://doi.org/10.5067/MODIS/MCD12Q1.006, Friedl and Sulla-Menashe, 2019; https://doi.org/10.5067/MODIS/MCD15A2H.006, Myneni et al., 2015; https://doi.org/10.5067/MODIS/MCD43A4.006, Schaaf and Wang, 2015; https://doi.org/10.5067/MODIS/MOD11A1.006, Wan et al., 2015). ERA-5 Land is available through the Copernicus Climate Change Service (https://doi.org/10.24381/CDS.68D2BB30, Muñoz Sabater, 2019). Data from BESS radiation can be downloaded from the Bess_Rad website (https://www.environment.snu.ac.kr/bess-rad, Ryu et al., 2022). CSIF data are available through the National Tibetan Plateau/Third Pole Environment Data Center (https://doi.org/10.11888/Ecolo.tpdc.271751, Zhang, 2021). ESA CCI can be downloaded from the CEDA Archive (https://data.ceda.ac.uk/neodc/esacci/soil_moisture, Dorigo et al., 2022). ALEXI ET data are available upon request from the authors (Hain and Anderson, 2017). The FluxCom dataset can be downloaded from the Max Planck Institute for Biogeochemistry (https://www.bgc-jena.mpg.de/geodb/projects/Home.php, Boenisch, 2020). FluxSat is available from the ORNL DAAC (https://doi.org/10.3334/ORNLDAAC/1835, Joiner and Yoshida, 2021).

Author contributions. The study was conceptualized by YK and MG. YK contributed to the data curation. MG performed the formal analysis and developed the experimental methodology. MG prepared the manuscript draft, with contributions from YK and the other co-authors. The project was supervised by TK and GS.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. We would like to express our gratitude to Martha Anderson and Christopher Hain for providing the ALEXI ET dataset, which has greatly enriched our research. Furthermore, we are grateful for the provision of the ERA-5 Land dataset, which was generated using Copernicus Climate Change Service information 2019, and for the AMERIFLUX data portal, which is funded by the U.S. Department of Energy Office of Science. We would also like to thank Jiangong Liu and the anonymous reviewers for their helpful contributions.

Financial support. This research has been supported by the Department of Energy, Labor and Economic Growth (grant no. DE-SC0021023) and the National Aeronautics and Space Administration (grant nos. 80NSSC21K1705 and 80NSSC20K1801).

Review statement. This paper was edited by Andrew Feldman and reviewed by Jiangong Liu and two anonymous referees.

References

Ahlström, A., Raupach, M. R., Schurgers, G., Smith, B., Arneth, A., Jung, M., Reichstein, M., Canadell, J. G., Friedlingstein, P., Jain, A. K., Kato, E., Poulter, B., Sitch, S., Stocker, B. D., Viovy, N., Wang, Y. P., Wiltshire, A., Zaehle, S., and Zeng, N.: The dominant role of semi-arid ecosystems in the trend and variability of the land CO_2 sink, Science, 348, 895–899, https://doi.org/10.1126/science.aaa1668, 2015.

AmeriFlux Management Project: AmeriFlux, AmeriFlux [data set], https://ameriflux.lbl.gov/data/flux-data-products, last access: 13 October 2022.

Anav, A., Friedlingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C., Murray-Tortarolo, G., Papale, D., Parazoo, N. C., Peylin, P., Piao, S., Sitch, S., Viovy, N., Wiltshire, A., and Zhao, M.: Spatiotemporal patterns of terrestrial gross primary production: A review: GPP Spatiotemporal Patterns, Rev. Geophys., 53, 785–818, https://doi.org/10.1002/2015RG000483, 2015.

Babaeian, E., Paheding, S., Siddique, N., Devabhaktuni, V. K., and Tuller, M.: Estimation of root zone soil moisture from ground and remotely sensed soil information with multisensor data fusion

and automated machine learning, Remote Sens. Environ., 260, 112434, https://doi.org/10.1016/j.rse.2021.112434, 2021.

Balaji, A. and Allen, A.: Benchmarking Automatic Machine Learning Frameworks, ArXiv [preprint], https://doi.org/10.48550/arXiv.1808.06492, 2018.

Barnes, M. L., Farella, M. M., Scott, R. L., Moore, D. J. P., Ponce-Campos, G. E., Biederman, J. A., MacBean, N., Litvak, M. E., and Breshears, D. D.: Improved dryland carbon flux predictions with explicit consideration of water-carbon coupling, Commun. Earth Environ., 2, 1–9, https://doi.org/10.1038/s43247-021-00308-2, 2021.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M. A., Baldocchi, D., Bonan, G. B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K. W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F. I., and Papale, D.: Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate, Science, 329, 834–838, https://doi.org/10.1126/science.1184984, 2010.

Bloomfield, K. J., Stocker, B. D., Keenan, T. F., and Prentice, I. C.: Environmental controls on the light use efficiency of terrestrial gross primary production, Glob. Change Biol., 29, 1037–1053, https://doi.org/10.1111/gcb.16511, 2023.

Bodesheim, P., Jung, M., Gans, F., Mahecha, M. D., and Reichstein, M.: Upscaled diurnal cycles of land–atmosphere fluxes: a new global half-hourly data product, Earth Syst. Sci. Data, 10, 1327–1365, https://doi.org/10.5194/essd-10-1327-2018, 2018.

Boenisch, G.: Max Planck Institute for Biogeochemistry, Data Portal, https://www.bgc-jena.mpg.de/geodb/projects/Home.php (last access: 11 May 2023), 2020.

Canadell, J. G., Scheel Monteiro, P., Costa, M. H., Cotrim da Cunha, L., Cox, P. M., Eliseev, A. V., Henson, S., Ishii, M., Jaccard, S., Koven, C., Lohila, A., Patra, P. K., Piao, S., Rogelj, J., Syampungani, S., Zaehle, S., and Zickfeld, K.: Global carbon and other biogeochemical cycles and feedbacks, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, Ö., Yu, R., and Zhou, B., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 673–816, https://doi.org/10.1017/9781009157896.001, 2021.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A.: Ensemble selection from libraries of models, in: Twenty-first international conference on Machine learning – ICML '04, Twenty-first international conference, 4–8 July 2004, Banff, Alberta, Canada, 18, https://doi.org/10.1145/1015330.1015432, 2004.

Carvalho, S., Oliveira, A., Pedersen, J. S., Manhice, H., Lisboa, F., Norguet, J., de Wit, F., and Santos, F. D.: A changing Amazon rainforest: Historical trends and future projections under post-Paris climate scenarios, Global Planet. Change, 195, 103328, https://doi.org/10.1016/j.gloplacha.2020.103328, 2020.

Cawley, G. C. and Talbot, N. L. C.: On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, J. Mach. Learn. Res., 11, 2079–2107, 2010.

Chen, C., Park, T., Wang, X., Piao, S., Xu, B., Chaturvedi, R. K., Fuchs, R., Brovkin, V., Ciais, P., Fensholt, R., Tømmervik, H., Bala, G., Zhu, Z., Nemani, R. R., and Myneni, R. B.: China and India lead in greening of the world through land-use management, Nat. Sustain., 2, 122–129, https://doi.org/10.1038/s41893-019-0220-7, 2019.

Chu, H., Luo, X., Ouyang, Z., Chan, W. S., Dengel, S., Biraud, S. C., Torn, M. S., Metzger, S., Kumar, J., Arain, M. A., Arkebauer, T. J., Baldocchi, D., Bernacchi, C., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Bracho, R., Brown, S., Brunsell, N. A., Chen, J., Chen, X., Clark, K., Desai, A. R., Duman, T., Durden, D., Fares, S., Forbrich, I., Gamon, J. A., Gough, C. M., Griffis, T., Helbig, M., Hollinger, D., Humphreys, E., Ikawa, H., Iwata, H., Ju, Y., Knowles, J. F., Knox, S. H., Kobayashi, H., Kolb, T., Law, B., Lee, X., Litvak, M., Liu, H., Munger, J. W., Noormets, A., Novick, K., Oberbauer, S. F., Oechel, W., Oikawa, P., Papuga, S. A., Pendall, E., Prajapati, P., Prueger, J., Quinton, W. L., Richardson, A. D., Russell, E. S., Scott, R. L., Starr, G., Staebler, R., Stoy, P. C., Stuart-Haëntjens, E., Sonnentag, O., Sullivan, R. C., Suyker, A., Ueyama, M., Vargas, R., Wood, J. D., and Zona, D.: Representativeness of Eddy-Covariance flux footprints for areas surrounding AmeriFlux sites, Agr. Forest Meteorol., 301–302, 108350, https://doi.org/10.1016/j.agrformet.2021.108350, 2021.

Dannenberg, M. P., Barnes, M. L., Smith, W. K., Johnston, M. R., Meerdink, S. K., Wang, X., Scott, R. L., and Biederman, J. A.: Upscaling dryland carbon and water fluxes with artificial neural networks of optical, thermal, and microwave satellite remote sensing, Biogeosciences, 20, 383–404, https://doi.org/10.5194/bg-20-383-2023, 2023.

Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets, J. Mach. Learn. Res., 7, 1–30, 2006.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S. I., Smolander, T., and Lecomte, P.: ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions, Remote Sens. Environ., 203, 185–215, https://doi.org/10.1016/j.rse.2017.07.001, 2017.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S. I., Smolander, T., and Lecomte, P.: Data from the ESA CCI Soil Moisture project, CEDA Archive [data set], https://data.ceda.ac.uk/neodc/esacci/soil_moisture, last access: 13 October 2022.

Duan, S. and Zhang, X.: AutoML-Based Drought Forecast with Meteorological Variables, arXiv [preprint], https://doi.org/10.48550/arXiv.2207.07012, 23 August 2022.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A.: AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data, arXiv [preprint], https://doi.org/10.48550/arXiv.2003.06505, 13 March 2020.

Ferreira, L., Pilastri, A., Martins, C. M., Pires, P. M., and Cortez, P.: A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost, in: 2021 International Joint Conference on Neural Networks (IJCNN),

2021 International Joint Conference on Neural Networks (IJCNN), 18–22 July 2021, Shenzhen, China, 1–8, https://doi.org/10.1109/IJCNN52387.2021.9534091, 2021.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F.: Efficient and Robust Automated Machine Learning, in: Advances in Neural Information Processing Systems, Conference on Neural Information Processing Systems, Montréal, Canada, 7–12 December 2015, 2015a.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F., Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R.: auto-sklearn, GitHub [code], https://github.com/automl/auto-sklearn (last access: 13 August 2022), 2015b.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., and Hutter, F.: Practical Automated Machine Learning for the AutoML Challenge 2018, The Thirty-fifth International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018, 2018.

Frank, D., Reichstein, M., Bahn, M., Thonicke, K., Frank, D., Mahecha, M. D., Smith, P., van der Velde, M., Vicca, S., Babst, F., Beer, C., Buchmann, N., Canadell, J. G., Ciais, P., Cramer, W., Ibrom, A., Miglietta, F., Poulter, B., Rammig, A., Seneviratne, S. I., Walz, A., Wattenbach, M., Zavala, M. A., and Zscheischler, J.: Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts, Glob. Change Biol., 21, 2861–2880, https://doi.org/10.1111/gcb.12916, 2015.

Friedl, M. and Sulla-Menashe, D.: MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006, NASA EOSDIS Land Processes DAAC [data set], https://doi.org/10.5067/MODIS/MCD12Q1.006, 2019.

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Hauck, J., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Bakker, D. C. E., Canadell, J. G., Ciais, P., Jackson, R. B., Anthoni, P., Barbero, L., Bastos, A., Bastrikov, V., Becker, M., Bopp, L., Buitenhuis, E., Chandra, N., Chevallier, F., Chini, L. P., Currie, K. I., Feely, R. A., Gehlen, M., Gilfillan, D., Gkritzalis, T., Goll, D. S., Gruber, N., Gutekunst, S., Harris, I., Haverd, V., Houghton, R. A., Hurtt, G., Ilyina, T., Jain, A. K., Joetzjer, E., Kaplan, J. O., Kato, E., Klein Goldewijk, K., Korsbakken, J. I., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lenton, A., Lienert, S., Lombardozzi, D., Marland, G., McGuire, P. C., Melton, J. R., Metzl, N., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-I., Neill, C., Omar, A. M., Ono, T., Peregon, A., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Séférian, R., Schwinger, J., Smith, N., Tans, P. P., Tian, H., Tilbrook, B., Tubiello, F. N., van der Werf, G. R., Wiltshire, A. J., and Zaehle, S.: Global Carbon Budget 2019, Earth Syst. Sci. Data, 11, 1783–1838, https://doi.org/10.5194/essd-11-1783-2019, 2019.

Gaber, M.: mxgbr/gpp_upscaling: AutoML for GPP upscaling v1.0 (publication), Zenodo [code], https://doi.org/10.5281/zenodo.8262618, 2023.

Gong, L. J., Liu, S. M., Shuang, X., Cai, X. H., and Xu, Z. W.: Investigation of spatial representativeness for surface flux measurements with eddy covariance system and large aperture scintillometer, Plateau Meteorology, 28, 246–257, 2009.

Gray, J. M., Frolking, S., Kort, E. A., Ray, D. K., Kucharik, C. J., Ramankutty, N., and Friedl, M. A.: Direct human influence on atmospheric CO_2 seasonality from increased cropland productivity, Nature, 515, 398–401, https://doi.org/10.1038/nature13957, 2014.

Green, J. K., Seneviratne, S. I., Berg, A. M., Findell, K. L., Hagemann, S., Lawrence, D. M., and Gentine, P.: Large influence of soil moisture on long-term terrestrial carbon uptake, Nature, 565, 476–479, https://doi.org/10.1038/s41586-018-0848-x, 2019.

Green, J. K., Ballantyne, A., Abramoff, R., Gentine, P., Makowski, D., and Ciais, P.: Surface temperatures reveal the patterns of vegetation water stress and their environmental drivers across the tropical Americas, Glob. Change Biol., 28, 2940–2955, https://doi.org/10.1111/gcb.16139, 2022.

Gruber, A., Scanlon, T., van der Schalie, R., Wagner, W., and Dorigo, W.: Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology, Earth Syst. Sci. Data, 11, 717–739, https://doi.org/10.5194/essd-11-717-2019, 2019.

Guevara-Escobar, A., González-Sosa, E., Cervantes-Jiménez, M., Suzán-Azpiri, H., Queijeiro-Bolaños, M. E., Carrillo-Ángeles, I., and Cambrón-Sandoval, V. H.: Machine learning estimates of eddy covariance carbon flux in a scrub in the Mexican highland, Biogeosciences, 18, 367–392, https://doi.org/10.5194/bg-18-367-2021, 2021.

Guyon, I., Sun-Hosoya, L., Boullé, M., Escalante, H. J., Escalera, S., Liu, Z., Jajetic, D., Ray, B., Saeed, M., Sebag, M., Statnikov, A., Tu, W.-W., and Viegas, E.: Analysis of the AutoML Challenge Series 2015–2018, in: Automated Machine Learning: Methods, Systems, Challenges, edited by: Hutter, F., Kotthoff, L., and Vanschoren, J., Springer International Publishing, Cham, 177–219, https://doi.org/10.1007/978-3-030-05318-5_10, 2019.

Hain, C. R. and Anderson, M. C.: Estimating morning change in land surface temperature from MODIS day/night observations: Applications for surface energy balance modeling, Geophys. Res. Lett., 44, 9723–9733, https://doi.org/10.1002/2017GL074952, 2017.

Hanussek, M., Blohm, M., and Kintz, M.: Can AutoML outperform humans? An evaluation on popular OpenML datasets using AutoML Benchmark, in: 2020 2nd International Conference on Artificial Intelligence, Robotics and Control, 12–14 December 2020, New York, NY, USA, 29–32, https://doi.org/10.1145/3448326.3448353, 2020.

Haughton, N., Abramowitz, G., De Kauwe, M. G., and Pitman, A. J.: Does predictability of fluxes vary between FLUXNET sites?, Biogeosciences, 15, 4495–4513, https://doi.org/10.5194/bg-15-4495-2018, 2018.

Hutter, F., Kotthoff, L., and Vanschoren, J. (Eds.): Automated Machine Learning: Methods, Systems, Challenges, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-05318-5, 2019.

International Geosphere–Biosphere Programme: IGBP, http://www.igbp.net, last access: 8 January 2024.

Joiner, J. and Yoshida, Y.: Satellite-based reflectances capture large fraction of variability in global gross primary production (GPP) at weekly time scales, Agr. Forest Meteorol., 291, 108092, https://doi.org/10.1016/j.agrformet.2020.108092, 2020.

Joiner, J. and Yoshida, Y.: Global MODIS and FLUXNET-derived Daily Gross Primary Production, V2 (2), ORNL Distributed Active Archive Center [data set], https://doi.org/10.3334/ORNLDAAC/1835, 2021.

Joiner, J., Yoshida, Y., Zhang, Y., Duveiller, G., Jung, M., Lyapustin, A., Wang, Y., and Tucker, C.: Estimation of Terrestrial Global Gross Primary Production (GPP) with Satellite Data-Driven Models and Eddy Covariance Flux Data, Remote Sens.-Basel, 10, 1346, https://doi.org/10.3390/rs10091346, 2018.

Jones, P. W.: First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates, Mon. Weather Rev., 127, 2204–2210, https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2, 1999.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, J. Geophys. Res., 116, G00J07, https://doi.org/10.1029/2010JG001566, 2011.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., Chevallier, F., Gans, F., Goll, D. S., Haverd, V., Köhler, P., Ichii, K., Jain, A. K., Liu, J., Lombardozzi, D., Nabel, J. E. M. S., Nelson, J. A., O'Sullivan, M., Pallandt, M., Papale, D., Peters, W., Pongratz, J., Rödenbeck, C., Sitch, S., Tramontana, G., Walker, A., Weber, U., and Reichstein, M.: Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach, Biogeosciences, 17, 1343–1365, https://doi.org/10.5194/bg-17-1343-2020, 2020.

Kalfas, J. L., Xiao, X., Vanegas, D. X., Verma, S. B., and Suyker, A. E.: Modeling gross primary production of irrigated and rain-fed maize using MODIS imagery and CO_2 flux tower data, Agr. Forest Meteorol., 151, 1514–1528, https://doi.org/10.1016/j.agrformet.2011.06.007, 2011.

Kannenberg, S. A., Anderegg, W. R. L., Barnes, M. L., Dannenberg, M. P., and Knapp, A. K.: Dominant role of soil moisture in mediating carbon and water fluxes in dryland ecosystems, Nat. Geosci., 17, 38–43, https://doi.org/10.1038/s41561-023-01351-8, 2024.

Karmaker, S. K., Hassan, M. M., Smith, M. J., Xu, L., Zhai, C., and Veeramachaneni, K.: AutoML to Date and Beyond: Challenges and Opportunities, ACM Comput. Surv., 54, 175:1–175:36, https://doi.org/10.1145/3470918, 2021.

Keenan, T. F., Prentice, I. C., Canadell, J. G., Williams, C. A., Wang, H., Raupach, M., and Collatz, G. J.: Recent pause in the growth rate of atmospheric CO_2 due to enhanced terrestrial carbon uptake, Nat. Commun., 7, 13428, https://doi.org/10.1038/ncomms13428, 2016.

Keenan, T. F., Migliavacca, M., Papale, D., Baldocchi, D., Reichstein, M., Torn, M., and Wutzler, T.: Widespread inhibition of daytime ecosystem respiration, Nat. Ecol. Evol., 3, 407–415, https://doi.org/10.1038/s41559-019-0809-2, 2019.

Kim, G. E., Steller, M., and Olson, S.: Modeling watershed nutrient concentrations with AutoML, in: Proceedings of the 10th International Conference on Climate Informatics, New York, NY, USA, 86–90, https://doi.org/10.1145/3429309.3429322, 2020.

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., and Leyton-Brown, K.: Auto-WEKA: Automatic Model Selection and Hyperparameter Optimization in WEKA, in: Automated Machine Learning: Methods, Systems, Challenges, edited by: Hutter, F., Kotthoff, L., and Vanschoren, J., Springer International Publishing, Cham, 81–95, https://doi.org/10.1007/978-3-030-05318-5_4, 2019.

LeDell, E. and Poirier, S.: H2O AutoML: Scalable Automatic Machine Learning, The Thirty-seventh International Conference on Machine Learning, Online, 12–18 July 2020, 2020.

Lee, S., Kim, J., Bae, J. H., Lee, G., Yang, D., Hong, J., and Lim, K. J.: Development of Multi-Inflow Prediction Ensemble Model Based on Auto-Sklearn Using Combined Approach: Case Study of Soyang River Dam, Hydrology, 10, 90, https://doi.org/10.3390/hydrology10040090, 2023.

Madni, H. A., Umer, M., Ishaq, A., Abuzinadah, N., Saidani, O., Alsubai, S., Hamdi, M., and Ashraf, I.: Water-Quality Prediction Based on H2O AutoML and Explainable AI Techniques, Water, 15, 475, https://doi.org/10.3390/w15030475, 2023.

Muñoz Sabater, J.: ERA5-Land monthly averaged data from 1950 to present, Copernicus Climate Change Service [data set], https://doi.org/10.24381/CDS.68D2BB30, 2019.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Myneni, R., Knyazikhin, Y., and Park, T.: MCD15A2H MODIS/Terra+Aqua Leaf Area Index/FPAR 8-day L4 Global 500m SIN Grid V006, NASA EOSDIS Land Processes Distributed Active Archive Center [data set], https://doi.org/10.5067/MODIS/MCD15A2H.006, 2015.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Orth, R., Destouni, G., Jung, M., and Reichstein, M.: Large-scale biospheric drought response intensifies linearly with drought duration in arid regions, Biogeosciences, 17, 2647–2656, https://doi.org/10.5194/bg-17-2647-2020, 2020.

Papale, D., Black, T. A., Carvalhais, N., Cescatti, A., Chen, J., Jung, M., Kiely, G., Lasslop, G., Mahecha, M. D., Margolis, H., Merbold, L., Montagnani, L., Moors, E., Olesen, J. E., Reichstein, M., Tramontana, G., Gorsel, E., Wohlfahrt, G., and Ráduly, B.: Effect of spatial sampling from European flux towers for estimating carbon and water fluxes with artificial neural networks, J. Geophys. Res.-Biogeo., 120, 1941–1957, https://doi.org/10.1002/2015JG002997, 2015.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E.,

da Rocha, H., Dai, X., Davis, K. J., Cinti, B. D., Grandcourt, A. de, Ligne, A. D., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., Tommasi, P. di, Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, Sci. Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., Cinti, B. D., Grandcourt, A. de, Ligne, A. D., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., Tommasi, P. di, Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., et al.: FLUXNET2015 dataset, FLUXNET Data Portal [data set], https://fluxnet.org/data/fluxnet2015-dataset, last access: 13 October 2022.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., and Pélissier, R.: Spatial validation reveals poor predictive performance of large-scale ecological mapping models, Nat. Commun., 11, 4540, https://doi.org/10.1038/s41467-020-18321-y, 2020.

Qi, W., Xu, C., and Xu, X.: AutoGluon: A revolutionary framework for landslide hazard analysis, Natural Hazards Research, 1, 103–108, https://doi.org/10.1016/j.nhres.2021.07.002, 2021.

Raschka, S.: Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, arXiv [preprint], https://doi.org/10.48550/arXiv.1811.12808, 10 November 2020.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grünwald, T., Havránková, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J.-M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D., and Valentini, R.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm, Glob. Change Biol., 11, 1424–1439, https://doi.org/10.1111/j.1365-2486.2005.001002.x, 2005.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, 2017.

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

Ryu, Y., Jiang, C., Kobayashi, H., and Detto, M.: MODIS-derived global land products of shortwave radiation and diffuse and total photosynthetically active radiation at 5km resolution from 2000, Remote Sens. Environ., 204, 812–825, https://doi.org/10.1016/j.rse.2017.09.021, 2018.

Ryu, Y., Jiang, C., Kobayashi, H., and Detto, M.: Bess_Rad, Bess_Rad Website [data set], https://www.environment.snu.ac.kr/bess-rad, last access: 13 October 2022.

Schaaf, C. and Wang, Z.: MCD43A4 MODIS/Terra+Aqua BRDF/Albedo Nadir BRDF Adjusted Ref Daily L3 Global – 500m V006, NASA EOSDIS Land Processes Distributed Active Archive Center [data set], https://doi.org/10.5067/MODIS/MCD43A4.006, 2015.

Schucknecht, A., Erasmi, S., Niemeyer, I., and Matschullat, J.: Assessing vegetation variability and trends in north-eastern Brazil using AVHRR and MODIS NDVI time series, Eur. J. Remote Sens., 46, 40–59, https://doi.org/10.5721/EuJRS20134603, 2013.

Smith, W. K., Biederman, J. A., Scott, R. L., Moore, D. J. P., He, M., Kimball, J. S., Yan, D., Hudson, A., Barnes, M. L., MacBean, N., Fox, A. M., and Litvak, M. E.: Chlorophyll Fluorescence Better Captures Seasonal and Interannual Gross Primary Productivity Dynamics Across Dryland Ecosystems of Southwestern North America, Geophys. Res. Lett., 45, 748–757, https://doi.org/10.1002/2017GL075922, 2018.

Smith, W. K., Dannenberg, M. P., Yan, D., Herrmann, S., Barnes, M. L., Barron-Gafford, G. A., Biederman, J. A., Ferrenberg, S., Fox, A. M., Hudson, A., Knowles, J. F., MacBean, N., Moore, D. J. P., Nagler, P. L., Reed, S. C., Rutherford, W. A., Scott, R. L., Wang, X., and Yang, J.: Remote sensing of dryland ecosystem structure and function: Progress, challenges, and opportunities, Remote Sens. Environ., 233, 111401, https://doi.org/10.1016/j.rse.2019.111401, 2019.

Stocker, B. D., Zscheischler, J., Keenan, T. F., Prentice, I. C., Peñuelas, J., and Seneviratne, S. I.: Quantifying soil moisture impacts on light use efficiency across biomes, New Phytol., 218, 1430–1449, https://doi.org/10.1111/nph.15123, 2018.

Stocker, B. D., Zscheischler, J., Keenan, T. F., Prentice, I. C., Seneviratne, S. I., and Peñuelas, J.: Drought impacts on terrestrial primary production underestimated by satellite monitoring, Nat. Geosci., 12, 264–270, https://doi.org/10.1038/s41561-019-0318-6, 2019.

Sulkava, M., Luyssaert, S., Zaehle, S., and Papale, D.: Assessing and improving the representativeness of monitoring networks: The European flux tower network example, J. Geophys. Res.-

Biogeo., 116, G00J04, https://doi.org/10.1029/2010JG001562, 2011.

Sun, W., Fang, Y., Luo, X., Shiga, Y. P., Zhang, Y., Andrews, A. E., Thoning, K. W., Fisher, J. B., Keenan, T. F., and Michalak, A. M.: Midwest US Croplands Determine Model Divergence in North American Carbon Fluxes, AGU Adv., 2, e2020AV000310, https://doi.org/10.1029/2020AV000310, 2021.

Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K.: Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 11–14 August 2013, Chicago, USA, 847–855, https://doi.org/10.1145/2487575.2487629, 2013.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences, 13, 4291–4313, https://doi.org/10.5194/bg-13-4291-2016, 2016.

Tramontana, G., Migliavacca, M., Jung, M., Reichstein, M., Keenan, T. F., Camps-Valls, G., Ogee, J., Verrelst, J., and Papale, D.: Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks, Glob. Change Biol., 26, 5235–5253, https://doi.org/10.1111/gcb.15203, 2020.

Traoré, K. R., Camero, A., and Zhu, X. X.: Compact Neural Architecture Search for Local Climate Zones Classification, in: 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, The 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), online, 6–8 October 2021, 393–398, https://doi.org/10.14428/esann/2021.ES2021-55, 2021.

Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, C. B., and Farivar, R.: Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools, in: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), 1471–1479, https://doi.org/10.1109/ICTAI.2019.00209, 2019.

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., and Ahl, D. E.: Evaluation of MODIS NPP and GPP products across multiple biomes, Remote Sens. Environ., 102, 282–292, https://doi.org/10.1016/j.rse.2006.02.017, 2006.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E.: Super Learner, Stat. Appl. Genet. Mo. B., 6, 25, https://doi.org/10.2202/1544-6115.1309, 2007.

von Buttlar, J., Zscheischler, J., Rammig, A., Sippel, S., Reichstein, M., Knohl, A., Jung, M., Menzer, O., Arain, M. A., Buchmann, N., Cescatti, A., Gianelle, D., Kiely, G., Law, B. E., Magliulo, V., Margolis, H., McCaughey, H., Merbold, L., Migliavacca, M., Montagnani, L., Oechel, W., Pavelka, M., Peichl, M., Rambal, S., Raschi, A., Scott, R. L., Vaccari, F. P., van Gorsel, E., Varlagin, A., Wohlfahrt, G., and Mahecha, M. D.: Impacts of droughts and extreme-temperature events on gross primary production and ecosystem respiration: a systematic assessment across ecosystems and climate zones, Biogeosciences, 15, 1293–1318, https://doi.org/10.5194/bg-15-1293-2018, 2018.

Walther, S., Besnard, S., Nelson, J. A., El-Madany, T. S., Migliavacca, M., Weber, U., Carvalhais, N., Ermida, S. L., Brümmer, C., Schrader, F., Prokushkin, A. S., Panov, A. V., and Jung, M.: Technical note: A view from space on global flux towers by MODIS and Landsat: the FluxnetEO data set, Biogeosciences, 19, 2805–2840, https://doi.org/10.5194/bg-19-2805-2022, 2022.

Wan, Z., Hook, S., and Hulley, G.: MOD11A1 MODIS-/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006, NASA EOSDIS Land Processes Distributed Active Archive Center [data set], https://doi.org/10.5067/MODIS/MOD11A1.006, 2015.

Wang, J., Feng, L., Palmer, P. I., Liu, Y., Fang, S., Bösch, H., O'Dell, C. W., Tang, X., Yang, D., Liu, L., and Xia, C.: Large Chinese land carbon sink estimated from atmospheric carbon dioxide data, Nature, 586, 720–723, https://doi.org/10.1038/s41586-020-2849-9, 2020.

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 (1.0), ICOS Carbon Portal [data set], https://doi.org/10.18160/2G60-ZHAK, 2022.

Wei, S., Yi, C., Fang, W., and Hendrey, G.: A global study of GPP focusing on light-use efficiency in a random forest regression model, Ecosphere, 8, e01724, https://doi.org/10.1002/ecs2.1724, 2017.

Wohlfahrt, G. and Galvagno, M.: Revisiting the choice of the driving temperature for eddy covariance CO_2 flux partitioning, Agr. Forest Meteorol., 237–238, 135–142, https://doi.org/10.1016/j.agrformet.2017.02.012, 2017.

Xiao, J., Zhuang, Q., Baldocchi, D. D., Law, B. E., Richardson, A. D., Chen, J., Oren, R., Starr, G., Noormets, A., Ma, S., Verma, S. B., Wharton, S., Wofsy, S. C., Bolstad, P. V., Burns, S. P., Cook, D. R., Curtis, P. S., Drake, B. G., Falk, M., Fischer, M. L., Foster, D. R., Gu, L., Hadley, J. L., Hollinger, D. Y., Katul, G. G., Litvak, M., Martin, T. A., Matamala, R., McNulty, S., Meyers, T. P., Monson, R. K., Munger, J. W., Oechel, W. C., Paw U, K. T., Schmid, H. P., Scott, R. L., Sun, G., Suyker, A. E., and Torn, M. S.: Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data, Agr. Forest Meteorol., 148, 1827–1847, https://doi.org/10.1016/j.agrformet.2008.06.015, 2008.

Xu, H., Xiao, J., and Zhang, Z.: Heatwave effects on gross primary production of northern mid-latitude ecosystems, Environ. Res. Lett., 15, 074027, https://doi.org/10.1088/1748-9326/ab8760, 2020.

Yan, D., Scott, R. L., Moore, D. J. P., Biederman, J. A., and Smith, W. K.: Understanding the relationship between vegetation greenness and productivity across dryland ecosystems through the integration of PhenoCam, satellite, and eddy covariance data, Remote Sens. Environ., 223, 50–62, https://doi.org/10.1016/j.rse.2018.12.029, 2019.

Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., Yang, Q., and Yu, Y.: Taking Human out of Learning Applications: A Survey on Automated Machine Learning, arXiv [preprint], https://doi.org/10.48550/arXiv.1810.13306, 16 December 2019.

Yu, T., Sun, R., Xiao, Z., Zhang, Q., Liu, G., Cui, T., and Wang, J.: Estimation of Global Vegetation Productivity from Global LAnd Surface Satellite Data, Remote Sens.-Basel, 10, 327, https://doi.org/10.3390/rs10020327, 2018.

Yuan, W., Cai, W., Xia, J., Chen, J., Liu, S., Dong, W., Merbold, L., Law, B., Arain, A., Beringer, J., Bernhofer, C., Black, A., Blanken, P. D., Cescatti, A., Chen, Y., Francois, L., Gianelle, D., Janssens, I. A., Jung, M., Kato, T., Kiely, G., Liu, D., Marcolla, B., Montagnani, L., Raschi, A., Roupsard, O., Varlagin, A., and Wohlfahrt, G.: Global comparison of light use efficiency models for simulating terrestrial vegetation gross primary production based on the LaThuile database, Agr. Forest Meteorol., 192–193, 108–120, https://doi.org/10.1016/j.agrformet.2014.03.007, 2014.

Zhang, Y.: A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks (2000–2022), National Tibetan Plateau/Third Pole Environment Data Center [data set], https://doi.org/10.11888/Ecolo.tpdc.271751, 2021.

Zhang, Y., Xiao, X., Zhou, S., Ciais, P., McCarthy, H., and Luo, Y.: Canopy and physiological controls of GPP during drought and heat wave, Geophys. Res. Lett., 43, 3325–3333, https://doi.org/10.1002/2016GL068501, 2016.

Zhang, Y., Joiner, J., Alemohammad, S. H., Zhou, S., and Gentine, P.: A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks, Biogeosciences, 15, 5779–5800, https://doi.org/10.5194/bg-15-5779-2018, 2018.

Zöller, M.-A. and Huber, M. F.: Benchmark and Survey of Automated Machine Learning Frameworks, arXiv [preprint], https://doi.org/10.48550/arXiv.1904.12054, 26 January 2021.