https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

Biogeosciences Discussions
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], 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 to: Max Gaber (mfg@ign.ku.dk), Yanghui Kang (yanghuikang@berkeley.edu), Trevor Keenan (trevorkeenan@berkeley.edu)

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 scales 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, parametrization, and choice of predictor features. 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 predictor variables. Predictor variables from only MODIS surface reflectance data and photosynthetically active radiation explained over 70% of the monthly variability in GPP, while satellite-derived proxies for land surface temperature, evapotranspiration, soil moisture and plant functional types, and climate variables from reanalysis (ERA5-Land) further improved the frameworks' predictive ability. We found that the AutoML framework AutoSklearn consistently outperformed other AutoML frameworks as well as a classical Random Forest regressor in predicting GPP, reaching an overall r^2 of 0.75. In addition, we deployed AutoSklearn 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 hundred 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

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

to a global scale. These 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). These methods use globally available explanatory data, such as from satellite measurements and climate datasets, to establish a functional relationship to the GPP measurements, which can then be used to predict GPP globally. 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 importance of different predictor variables, such as greenness measures, photosynthetically active radiation (PAR), land surface temperature (LST), soil moisture (SM), and meteorological variables (vapor pressure deficit, temperature, precipitation) in controlling GPP 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, 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 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 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 among remote sensing and meteorological variables, the importance of different predictor 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.

Automated machine learning (AutoML) aims to overcome the challenges of algorithm selection, hyperparameter tuning, and pipeline creation through an automated approach. By evaluating different combinations of data processing steps, candidate ML models, and hyperparameters, AutoML aims to find the ML pipeline configuration best suited 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 is 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 from *in situ* measurements using globally available explanatory variables. Specifically, we benchmark three different frameworks, AutoSklearn, H2O AutoML, and AutoGluon, which 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 ML algorithms and pipelines based on site-level measurements. In addition, we evaluate the predictive contribution of various remotely sensed vegetation structure variables (i.e., greenness, land surface temperature, soil moisture, evapotranspiration) and meteorological factors within the AutoML

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

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 spatio-temporal 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/), and ICOS Warm Winter 2020 (ICOS, 2020) 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 night-time partitioning approach (Reichstein et al., 2005), and negative GPP outliers were truncated at -1 gC m^{-2} d^{-1}.

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): croplands (CRO), shrublands (SH), deciduous broadleaf forests (DBF), evergreen broadleaf forests (EBF), evergreen needleleaf forests (ENF), grasslands (GRA), mixed forests (MF), savannas (SAV), permanent wetlands (WET), and the non-vegetated classes of permanent snow and ice (SNO), water bodies (WAT), and barren soil (BAR).

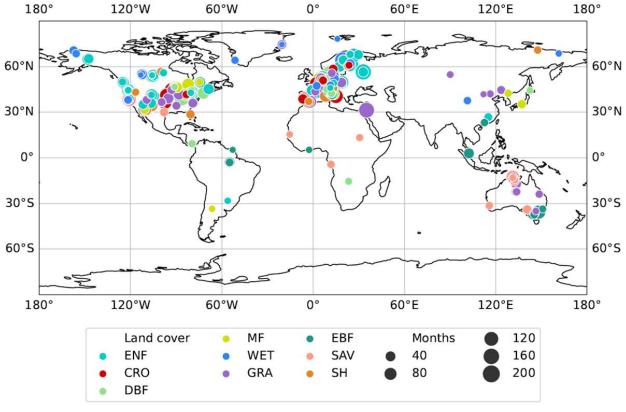

Figure 1: Locations of the measurement sites. The marker size represents the number of monthly measurements available at the respective location. The color stands for the land cover class reported at the site and comprises croplands (CRO), shrublands (SH), deciduous broadleaf forests (DBF), evergreen broadleaf forests (EBF), evergreen needleleaf forests (ENF), grasslands (GRA), mixed forests (MF), savannas (SAV), and permanent wetlands (WET).

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

Biogeosciences Discussions Open Access EGU

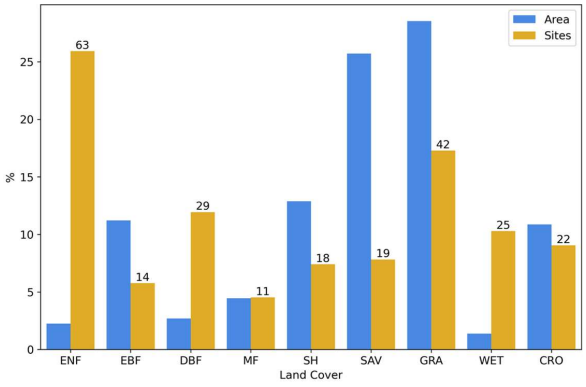

Figure 2: Standardized number of sites and global area of each land cover type, excluding land covers without any GPP measurements. The number of sites is shown above their respective columns. The land cover class reported comprises croplands (CRO), shrublands (SH), deciduous broadleaf forests (DBF), evergreen broadleaf forests (EBF), evergreen needleleaf forests (ENF), grasslands (GRA), mixed forests (MF), savannas (SAV), and permanent wetlands (WET).

2.1.2 Explanatory variables

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-BRDF adjusted reflectances (NBAR) from optical to infrared wavelengths, several derived vegetation indices (VI), 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, and soil moisture. In addition, we used meteorological data from the ERA5-Land reanalysis, including precipitation, temperature, and vapor pressure deficit (VPD). Table 1 shows an overview of all predictor variables. We furthermore applied a three-month lag in precipitation to account for water availability. All datasets were resampled into a 0.05° spatial resolution.

Table 1: Predictor variables and sources and their respective spatial and temporal resolution. The vegetation indices are abbreviated with: NDVI (Normalized difference vegetation index), EVI (Enhanced vegetation index), GCI (Green chlorophyll index), NDWI (Normalized difference water index), NIRv (Near-infrared reflectance of vegetation), and kNDVI (Kernel NDVI).

Predictor Variable	Source	Spatial Resolution	Temporal Resolution
Reflectance (Nadir-BRDF adjusted; NBAR) Bands 1–7	MODIS MCD43C4 v006 (Schaaf and Wang, 2015)	0.05 °	daily
Vegetation indices (NDVI, EVI, GCI, NDWI, NIRv, kNDVI)	Based on MODIS MCD43C4 v006	0.05 °	
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)	500m	4 days
LAI	MODIS MCD15A2H v006 (Myneni et al., 2015)	500m	4 days
Land surface temperature (day)	MODIS MYD11A1, MOD11A1 (Wan et al., 2015)	1km	daily

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

Land surface temperature (night)	MODIS MYD11A1, MOD11A1 (Wan et al., 2015)	1km	daily
Evapotranspiration	ALEXI (Hain and Anderson, 2017)	0.05 °	daily
Soil moisture	ESA CCI (Gruber et al., 2019)	0.25 °	daily
SIF	CSIF (Zhang et al., 2018)	0.05 °	4 days
Instantaneous SIF	CSIF (Zhang et al., 2018)	0.05 °	4 days
Land cover (biome)	MODIS MCD12Q1 (Friedl and Sulla-Menashe, 2019)	500m	annual
Total precipitation	ERA5-Land (Muñoz-Sabater et al., 2021)	0.1 °	hourly
Total precipitation (3 months lag)	ERA5-and (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

2.2 Automated machine learning

The performance of ML is highly dependent on the selection and configuration of data processing steps, algorithms, 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 explanatory input features into a target variable (Zöller and Huber, 2021). 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 data processing methods and selects or combines the most appropriate candidates for the ML problem.

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, AutoSklearn, 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).

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

AutoSklearn

AutoSklearn (Feurer et al., 2015) is an AutoML library built on top of the Scikit-Learn ML models. We used AutoSklearn in version 0.14.7. The framework relies on a wide range of base models, including AdaBoost, ARD regression, Decision Trees, Extra Trees, Gaussian processes, Gradient Boosting, k-Nearest Neighbors, Support Vector regression, MLP regression, Random Forests, and SGD regression. It also considers feature engineering algorithms, such as PCA, percentile regression, and feature agglomeration (AutoSklearn, n.d.). The framework selects and tunes its base models in a Bayesian optimization and performs a forward stepwise ensemble selection (van der Laan et al., 2007). It 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 et al., 2015).

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 trains multiple base ML models and ensembles thereof by combining their predictions through a generalized linear model (GLM) by default. The framework trains the base models in a predefined order with increasing diversity and complexity, including the so-called model families of Gradient Boosting models (GBM), XGBoost GBMs, GLMs, a default Random Forest model (DRF), Extremely Randomized Trees (XRT), and feed-forward neural networks. These models are either trained with pre-specified default hyperparameters or tuned by random search. H2O AutoML selects the best-performing models from all trained base models and evaluates ensembles of all trained models and the best models from each model family. An internal cross-validation (CV) trains and ranks the different candidate models.

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 ensembled and stacked in multiple layers in a bootstrap aggregation (bagging) process with skip connections between the layers. Each layer consists of the same models and parameters as the base layer. It receives the prediction from the previous layer along with the original data features themselves as input and predicts the dependent variable that is input to the next layer. The final stacking layer contains a weighted ensemble selection (van der Laan et al., 2007). Furthermore, 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 four 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 AutoSklearn with the best-performing set of explanatory variables to upscale *in situ* eddy covariance GPP measurements to global wall-to-wall maps (Fig. 3).

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

Biogeosciences Discussions Open Access EGU

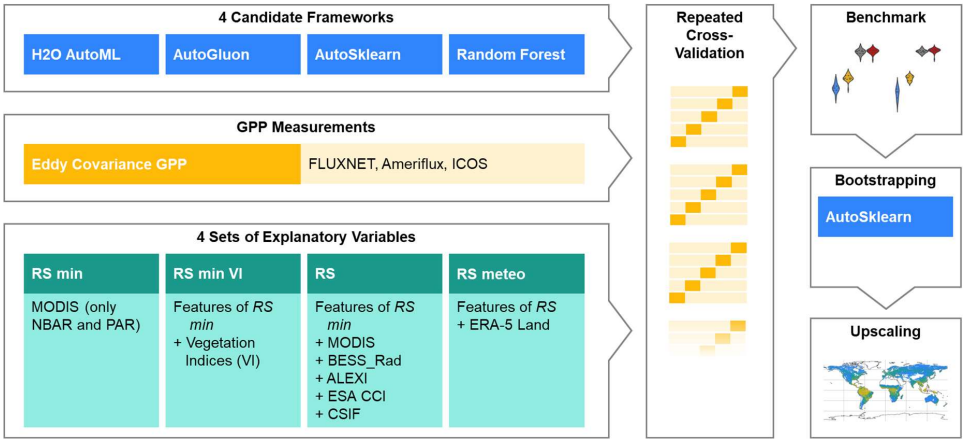

Figure 3 Experiment setup. We trained and evaluated AutoSklearn, H2O AutoML, AutoGluon, and Random Forest, together with four sets of explanatory variables in repeated cross-validation on GPP data from eddy covariance measurements. Then, we trained AutoSklearn in a bootstrap aggregation to produce global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

2.3.1 Explanatory variable sets

We organized the explanatory variables into four 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 next set of variables (RS minimal + VI) included additional VIs derived from MODIS bands which are designed to optimize the sensitivity to changes in vegetation structure and are widely used in predicting GPP and other ecological variables. The "RS" set included all remotely sensed variables and their products, except for VIs. 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.

Table 2 Predictor variable sets and associated data sets.

Predictor Variable	RS minimal	RS minimal +VI	RS	RS meteo
Reflectance (Nadir-BRDF adjusted; NBAR), Bands 1–7	●	●	●	●
Vegetation indices (NDVI, EVI, GCI, NDWI, NIRv, kNDVI)		●		
PAR	●	●	●	●
Diffuse PAR			●	●
RSDN			●	●
FPAR			●	●
LAI			●	●
Land surface temperature (day)			●	●
Land surface temperature (night)			●	●

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

ET	•	•
Soil moisture	•	•
SIF	•	•
Instantaneous SIF	•	•
Land cover (biome)	•	•
Total precipitation		•
Total precipitation (3 months lag)		•
Temperature		•
Vapor Pressure Deficit (VPD)		•

The explanatory variable sets can provide information about the importance of the input features to the upscaling framework. 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 five-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 thirty times with different random splits to evaluate the impact of partitioning the data on the final performance in our evaluation.

With H2O AutoML, AutoSklearn, 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 minutes on 20 CPUs) and 64GB 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. 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, seasonality (mean seasonal cycle) by month-wise 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. Only time series with more than 24 months of measurements were included in the evaluation.

Furthermore, 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) (Demšar, 2006).

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

2.3.3 GPP upscaling

We used AutoSklearn with the "RS" explanatory variable set to upscale the eddy covariance measurements to global scale, as this combination of framework and explanatory variables performed best in the benchmark. We trained thirty predictors 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 (Jung et al., 2020) and FluxSat (Joiner and Yoshida, 2020).

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 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 AutoSklearn performs best, followed by H2O AutoML, Random Forest, and AutoGluon in predicting monthly GPP (Fig. 4). AutoSklearn 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.

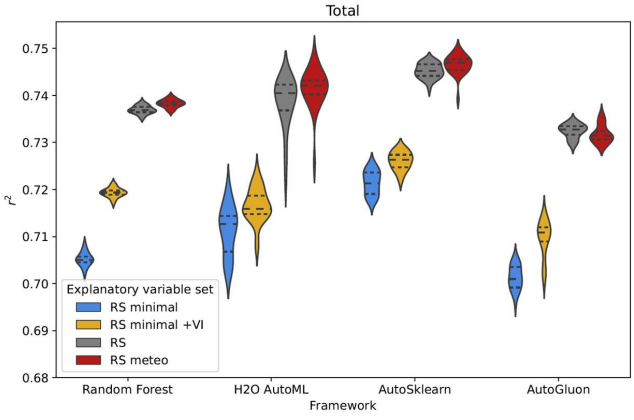

Figure 4: Overall framework performance, given in the coefficient of determination r^2 for the candidate frameworks and the four different predictor variable sets. Each distribution results from the repeated CV and belongs to one framework and one predictor variable set.

AutoSklearn's superior performance is primarily due to its ability to capture seasonal components, across-site variability, and trends (Fig. 5). When trained on "RS" explanatory variables, AutoSklearn achieved average r^2 values of 0.7452 ± 0.0003 overall, and 0.477 ± 0.003 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 (AutoSklearn: 10.31 ± 0.04%). Uncertainties are reported as the standard error of the mean of all cross-validation results.

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

Biogeosciences Discussions Open Access EGU

270

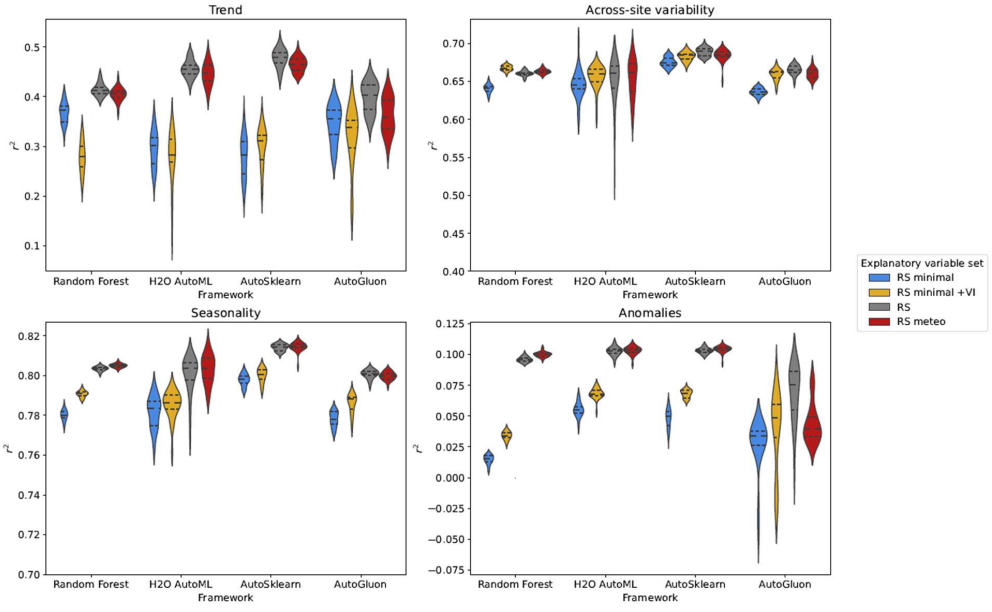

Figure 5: Evaluation of the temporally and spatially decomposed time series given in the coefficient of determination r^2. Each distribution results from the repeated CVs and belongs to one framework and one predictor variable set.

Using the Friedman test, we found that the four ML frameworks are statistically different in their performance in predicting
275 monthly GPP as well as its trends, seasonality, anomaly, and across-site variability (p-value < 0.01). The Nemenyi post hoc test shows that for the "RS" predictor variables, AutoSklearn achieves the highest average rank with statistical significance among all frameworks for monthly GPP and all its components (Fig. 6a). For the prediction of trends and anomalies, we could not find a significant difference in the average rank between AutoSklearn and H2O AutoML. Random Forest and AutoGluon perform the worst, while they are not statistically different in predicting trend, across site variability, and seasonal cycles.

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

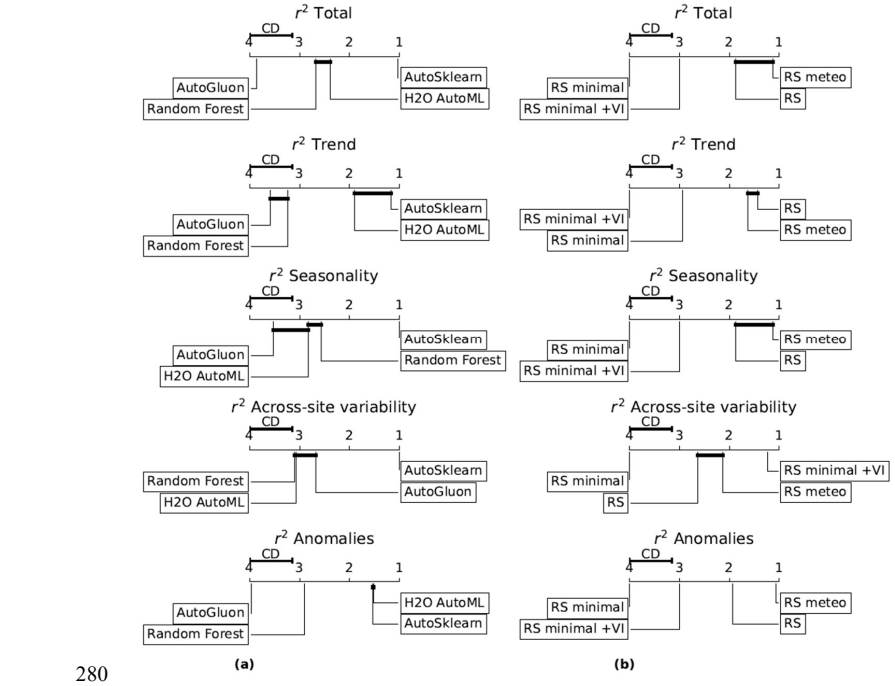

Figure 6 Critical difference (CD) diagrams (Demšar, 2006) for the ranks of the AutoML frameworks, trained on the "RS" predictor variables (a) and for the different predictor variable sets applied in the AutoSklearn framework trained on different predictor variable sets (b). Shown is the average rank distribution of the frameworks across all random cross-validations. A significant difference (CD) is obtained from a Nemenyi post hoc test, and rank differences smaller than the CD are marked with a connecting bar.

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). VIs provided additional predictive power in explaining across-site variability, seasonality, and anomalies, leading to a marginal increase in r^2 for monthly GPP. The greatest improvement occurred with the "RS" set when information on LST, ET, soil moisture, and biome type was included. The "RS" set increased r^2 on "RS minimal+VI" by about 0.02 for all frameworks, with sizable improvements in predicting trends and anomalies. Meteorological variables slightly improved the prediction of monthly GPP by better explaining spatial variability, seasonal cycle, and anomalies but reduced the models' ability to reproduce trends in all frameworks except AutoGluon. 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" statistically significantly for predicting GPP and all of its spatiotemporal components. Except for the performance of Random Forest on the across-site variability, "RS" was always the best-performing variable set or insignificantly different from the best-performing variable set.

Furthermore, we grouped the predictions by site and evaluated the site-level r^2 for each land cover type for AutoSklearn with "RS" explanatory variables (Fig. 7). 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.

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

Biogeosciences Discussions Open Access EGU

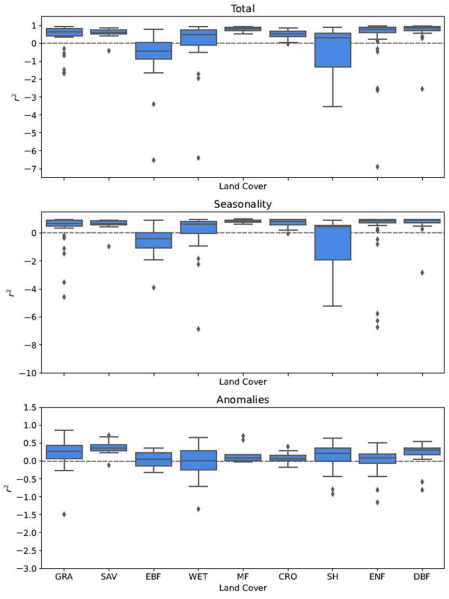

Figure 7 Distribution of r^2 values for the GPP prediction by AutoSklearn with "RS" explanatory variables for different land cover types. Shown are the overall performance and performances for seasonality and the anomalies.

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 AutoSklearn with the "RS" variable set. The prediction r^2 was with 0.8164 ± 0.0005 overall and 0.534 ± 0.003, 0.787 ± 0.002, 0.8723 ± 0.0005, and 0.3091 ± 0.0006 for trend, across-site variability, seasonality, and anomalies, respectively, significantly higher than for the lower resolution data product (Fig. 8).

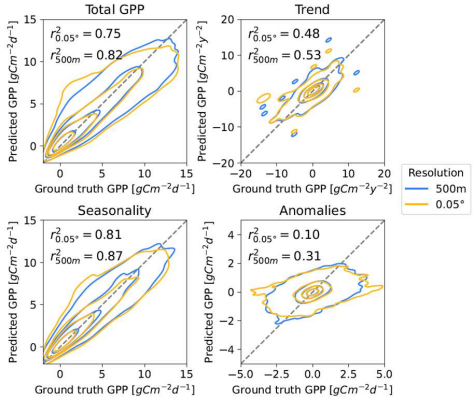

Figure 8 Comparison of the predicted 0.05 ° product and the one with 500m resolution from AutoSklearn ensemble averages and the "RS" variable set. The latter shows higher r^2 values compared to the ground truth GPP estimates from FLUXNET, AmeriFlux OneFlux, and ICOS. We refer to GPP measurements derived from eddy covariance at the flux tower locations as ground truth.

3.2 Analysis of AutoSklearn Pipelines

We investigated the different components (base models and preprocessing algorithms) of the AutoSklearn framework, which was trained on the "RS" variable set in the repeated cross-validation (See figure A1 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 zero to one. 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

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

Biogeosciences Discussions Open Access EGU

type, resulting in a distribution of scaled RMSEs for each base-model type (Fig. 9). 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 AutoSklearn. The ensemble selection algorithm (forward stepwise model selection) in AutoSklearn, 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 AutoSklearn.

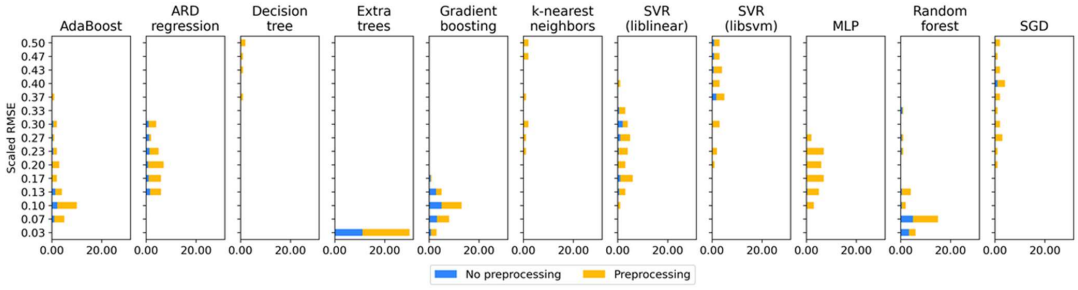

Figure 9 Performance of AutoSklearn base models and feature pre-processors. The chart shows the distribution of the mean RMSE for each base model type across all folds within each repetition of the cross-validation. We considered only the best-performing models for each model class within each fold. The RMSE is min-max scaled from zero to one 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 Figure A2).

3.3 Global GPP maps

From the bootstrap aggregation of the AutoSklearn 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).

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

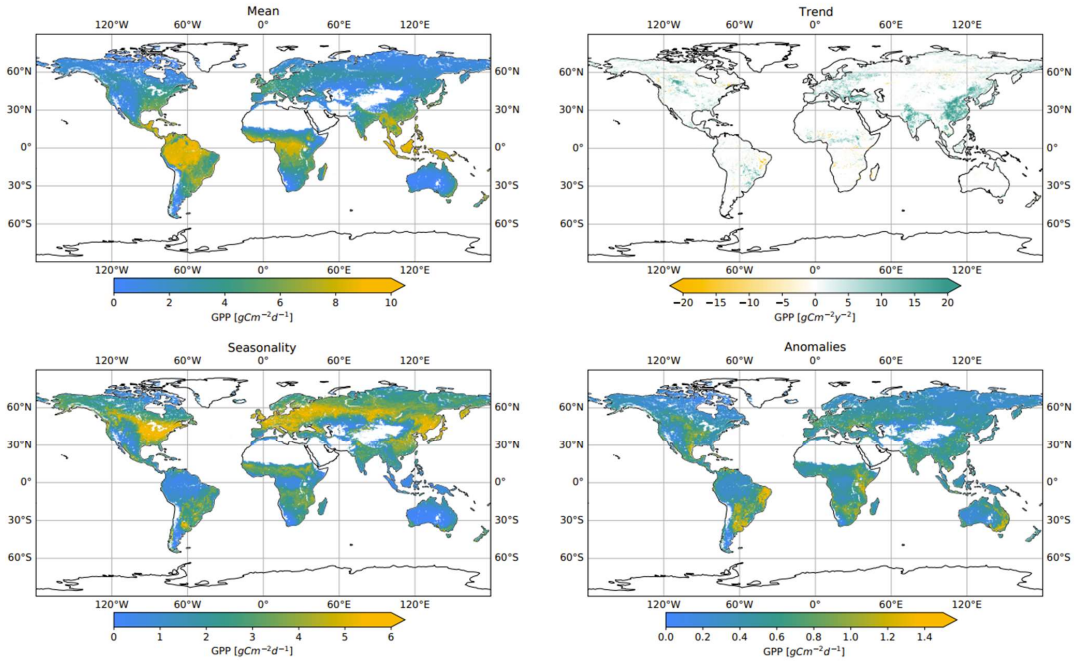

Figure 10 Total GPP, amplitude of seasonality, trend, and anomalies of prediction with AutoSklearn trained on remotely sensed data ("RS" dataset) in a bootstrap aggregation of 30 bootstraps. The mean was calculated at each location over all bootstrapped predictions and the entire time series. The seasonality is displayed as the amplitude of the month-wise 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.

Mean GPP for 2001–2020 (Fig. 10) 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. 10). 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 north-eastern 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 AutoSklearn 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. 11). 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 zero and ends in a long tail towards higher values for all biomes (Fig. 12a). 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 %).

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

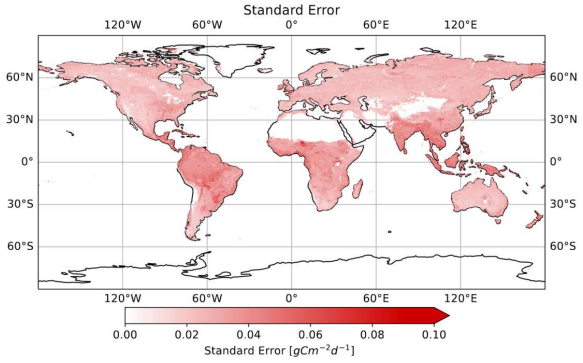

Figure 11 Absolute standard errors from the bootstrap aggregation. For relative values, see figure A3.

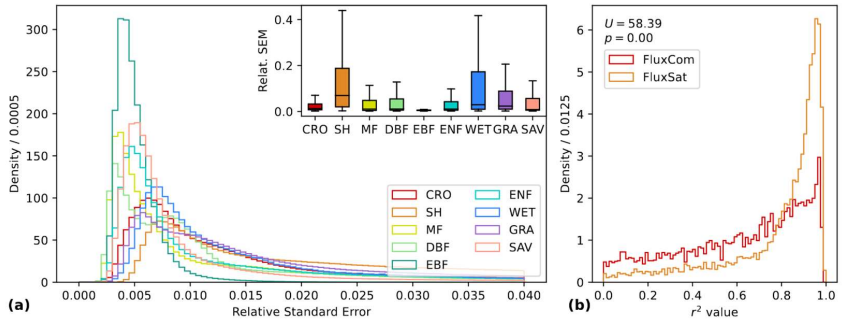

Figure 12 Histogram of the relative standard error of the mean (SEM) by land cover class during the entire observation period (a) and distribution of r^2 values for total GPP of the upscaled GPP AutoSklearn product with "RS" variables, compared to the FluxCom v6 and FluxSat datasets (b). For the latter, GPP is sampled at 10000 random locations and compared in a Mann-Whitney U test.

3.4 Comparison to reference data

We compared the upscaled results of total GPP from our AutoSklearn "RS" prediction with GPP datasets FluxCom v6 (Tramontana et al., 2016) and FluxSat (Joiner et al., 2018) at 10000 random sample locations. When tested with a Mann-Whitney U test, our predictions show significantly higher agreement (p virtually zero) with FluxSat than with FluxCom (Fig. 12b). In our prediction, 51 % of the samples explain more than 80 %, respectively, of the variation in FluxSat, while this is the case for only 17 % of the samples in FluxCom.

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).

However, there are significant performance differences between the frameworks. AutoSklearn significantly and 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 AutoSklearn outlines the applicability of various ML model types for predicting GPP. It is evident that ensembles of weak learners, such as Extra Trees

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

or Random Forest, are generally favorable for this task. These models can be promising for GPP prediction either in a stand-alone implementation or as part of a model ensemble. The outperformance against H2O AutoML and AutoGluon shows furthermore that the implementation of 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.

The differing prediction quality of AutoSklearn at land cover level might be affected by 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 evaluation, such as for SH, which includes both arid and subarctic shrublands. Finally, it is crucial to note that the metric only expresses how well the framework can reproduce measurements from the sample, which are limited in underrepresented areas. This circumstance entails that sites are repeatedly selected for validation during the repeated CV, which can inflate the performance metric and reduce variance.

While we grouped data by site and applied a land cover stratification during the CV to increase independence between the folds, we could not 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).

4.2 Importance of explanatory variables

AutoML is a powerful approach for assessing variable importance 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 feature importance, especially as features in GPP predicting often exhibit severe intercorrelations.

The frameworks' performance depends heavily on the selection of predictive features on which they are trained. The results show that only considering the seven NBAR bands and PAR from the "RS minimal" variable set does not provide the model with all the information necessary for a GPP prediction, and neither does adding vegetation indices. However, the complete set of "RS" variables contains additional information that all the frameworks can exploit. The additional variables in the "RS" variable set, such as evapotranspiration, land surface temperature, soil moisture, and plant function type, seem to include crucial environmental forcings that cannot be accounted for by only considering the variables on vegetation structure and radiation in "RS minimal" and "RS minimal +VI" (Green et al., 2019; Stocker et al., 2019; Xu et al., 2020). For example, environmental stress, such as heat waves and droughts, often cause 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 VIs and surface reflectance may not sufficiently capture the variability of GPP.

Including the meteorological predictors 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 machine learning frameworks can effectively use to predict GPP. A possible explanation is that the "RS" set already includes variables, such as LST, ET,

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

and soil moisture, that encode information about the instantaneous environmental stress on LUE or GPP due to unfavorable meteorological conditions. Additionally, the quality of the reanalyzed meteorological data may not sufficiently inform the machine learning models due to the presence of large uncertainties. For example, Joiner and Yoshida (2020) showed that using site-measured meteorological data rather than reanalyzed data significantly improved the performance of GPP predictions.

We found that besides selecting an appropriate set of explanatory variables, the resolution of the data highly affects prediction outcomes. Including 500m resolution data should reduce the mixed pixel problem and match the flux towers' footprints better with the pixel size of the gridded data sets. This led to improvements in all time series components, with exceptional increases in r^2 for the estimation of anomalies. It is to be explored how the representation can be improved even further, e.g., through better representing the flux tower footprints (Xiao et al., 2008; Yu et al., 2018; Chu et al., 2021).

4.3 Spatio-temporal patterns

The globally upscaled measurements could capture the variation of 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 compared 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). Furthermore, land cover change patterns from reference deforestation datasets could not be noticeably replicated in the GPP trends (Hansen et al., 2013). 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 in these cultivations, 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 semi-arid and temperate climates. 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. Semi-arid regions dominate the interannual variability of the global terrestrial carbon sink (Ahlström et al., 2015).

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. 12a) show similarities to the spread of r^2 values obtained from the framework benchmark (Fig. 7).

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

Higher standard errors can potentially be caused by the lack of representative measurement sites or the substantial geographical spread between sites in the training set. Another reason could be that some ecosystems are more predictable than others, given the current set of explanatory variables. The robustness of the prediction can depend on many factors, such as seasonal variability, species diversity, or spatial and temporal heterogeneity. For example, GPP can be predicted relatively well for ecosystems with a high proportion of biomass, such as broadleaf forests and mixed forests, arguably as they exhibit distinct seasonal and inter-annual vegetation growth patterns and productivity. Yet also, savannas and croplands show low relative uncertainties. In contrast, predicting GPP for sparse or heterogeneous vegetation cover, such as shrublands and wetlands, could be more challenging due to their complex biophysical and environmental characteristics.

The results delineate that AutoSklearn could not reliably infer a robust functional relationship in low-productivity regions, where it shows a significant positive bias. It is to further research to improve the performance in low-GPP regions. A method that could potentially enhance the prediction can be manually including dummy measurement sites in the masked regions. These sites would constantly report zero GPP and might improve estimates in similar regions, such as arid zones or seasonally snow-covered areas, which are the ones that are also less proportionately represented in the flux tower networks.

5 Conclusion

We investigated whether and how AutoML frameworks can improve global GPP upscaling from *in situ* measurements using AutoSklearn, 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.

AutoSklearn significantly and consistently outperformed the other frameworks across all sets of predictor 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.

We found that remotely sensed (RS) explanatory variables provided the best results in combination with the investigated frameworks. However, relying only on the MODIS NBAR reflectance bands, PAR, and vegetation indices ("RS minimal" and "RS minimal +VI") did not provide the models with sufficient information for GPP prediction without considering factors that indicated environmental stresses, such as evapotranspiration, land surface temperature, or soil moisture. On the other hand, 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 (AutoSklearn with "RS" predictor 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 predictions. 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.

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

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. By leveraging the power of AutoML, researchers and practitioners can minimize human intervention and subjectivity, leading to more robust and accurate GPP predictions. 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. Researchers must interpret and validate the results obtained through AutoML, ensuring that the chosen model and features align with ecological knowledge and scientific understanding. In this way, the integration of AutoML into GPP prediction can accelerate research and facilitate more informed decision-making in areas such as climate change mitigation and ecosystem management.

Appendix

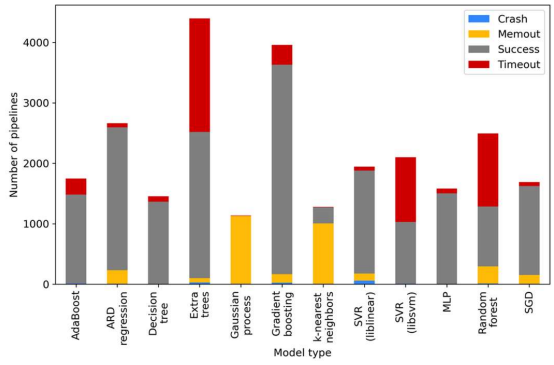

Figure A1 Run statistics of the AutoSklearn base models. The four statuses show how many base models succeeded or failed during training due to insufficient memory, training time, or other unknown reasons. Only the successful models were used for the configuration of AutoSklearn.

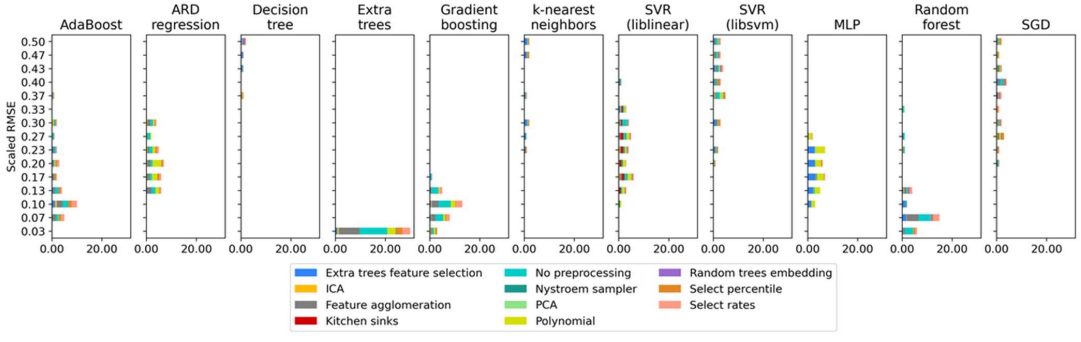

Figure A2 Detailed use of preprocessing algorithms by AutoSklearn.

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

Biogeosciences Discussions Open Access EGU

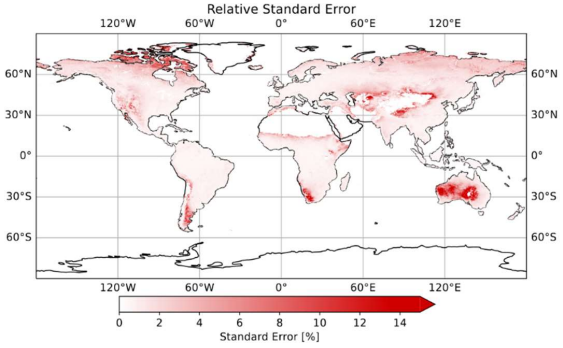

Figure A3 Relative average standard error, normalized by the mean GPP prediction.

Code availability

The code can be found at 10.5281/zenodo.8262618.

Author contribution

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 authors declare that they have no conflict of interest.

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. TK acknowledges funding from the LEMONTREE (Land Ecosystem Models based On New Theory, obseRvations and ExperimEnts) project, funded through the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures programme, a DOE Early Career Research Program award #DE-SC0021023, and NASA Awards 80NSSC21K1705 and 80NSSC20K1801. YK acknowledges support from a DOE Early Career Research Program award #DE-SC0021023 and the LEMONTREE project.

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.

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.

auto-sklearn/autosklearn/pipeline/components/feature_preprocessing at master · automl/auto-sklearn: https://github.com/automl/auto-sklearn, last access: 13 August 2022.

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

560 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 Sensing of Environment, 260, 112434, https://doi.org/10.1016/j.rse.2021.112434, 2021.

Balaji, A. and Allen, A.: Benchmarking Automatic Machine Learning Frameworks, ArXiv, 2018.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M. A., Baldocchi, D., Bonan,
565 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, https://doi.org/10.1126/science.1184984, 2010.

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

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-
575 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
580 international conference on Machine learning - ICML '04, Twenty-first international conference, 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 and Planetary Change, 195, 103328, https://doi.org/10.1016/j.gloplacha.2020.103328, 2020.

585 Cawley, G. C. and Talbot, N. L. C.: On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, 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.

590 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.,
595 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, Agricultural and Forest Meteorology, 301–302, 108350, https://doi.org/10.1016/j.agrformet.2021.108350, 2021.

600 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, Journal of Machine Learning Research, 7, 1–30, 2006.

605 Duan, S. and Zhang, X.: AutoML-Based Drought Forecast with Meteorological Variables, 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:2003.06505 [cs, stat], 2020.

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

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), 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, 9, 2015.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., and Hutter, F.: Practical Automated Machine Learning for the AutoML Challenge 2018, 2018.

Friedl, M. and Sulla-Menashe, D.: MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006, 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 System Science Data, 11, 1783–1838, https://doi.org/10.5194/essd-11-1783-2019, 2019.

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 CO2 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, Global Change Biology, 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 System Science 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, Geophysical Research Letters, 44, 9723–9733, https://doi.org/10.1002/2017GL074952, 2017.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., and Townshend, J. R. G.: High-Resolution Global Maps of 21st-Century Forest Cover Change, Science, 342, 850–853, https://doi.org/10.1126/science.1244693, 2013.

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, New York, NY, USA, 29–32, https://doi.org/10.1145/3448326.3448353, 2020.

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

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.

ICOS: Warm Winter 2020 ecosystem eddy covariance flux product for 73 stations in FLUXNET-Archive format—release 2022-1, 2020.

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

Joiner, J. and Yoshida, Y.: Vegetation CollectionGlobal MODIS and FLUXNET-derived Daily Gross Primary Production, V2, 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 Sensing, 10, 1346, https://doi.org/10.3390/rs10091346, 2018.

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 CO2 flux tower data, Agricultural and Forest Meteorology, 151, 1514–1528, https://doi.org/10.1016/j.agrformet.2011.06.007, 2011.

Karmaker, S. K., Hassan, Md. 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 CO2 due to enhanced terrestrial carbon uptake, Nat Commun, 7, 13428, https://doi.org/10.1038/ncomms13428, 2016.

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.

van der Laan, M. J., Polley, Eric C., and Hubbard, A. E.: Super Learner, U.C. Berkeley Division of Biostatistics, 2007.

LeDell, E. and Poirier, S.: H2O AutoML: Scalable Automatic Machine Learning, 7th ICML Workshop on Automated Machine Learning (AutoML), 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., 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.,

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

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, https://doi.org/10.5067/MODIS/MCD15A2H.006, 2015.

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. Biogeosci., 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.

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, 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, Global Change Biology, 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.

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 Sensing of Environment, 204, 812–825, https://doi.org/10.1016/j.rse.2017.09.021, 2018.

Schaaf, C. and Wang, Z.: MCD43A4 MODIS/Terra+Aqua BRDF/Albedo Nadir BRDF Adjusted Ref Daily L3 Global - 500m V006, 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, European Journal of Remote Sensing, 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, Geophysical Research Letters, 45, 748–757, https://doi.org/10.1002/2017GL075922, 2018.

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 Phytologist, 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, Journal of Geophysical Research: Biogeosciences, 116, 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 Advances, 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, New York, NY, 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 withregression algorithms, Biogeosciences, 13, 4291–4313, https://doi.org/10.5194/bg-13-4291-2016, 2016.

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, 393–398, 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), arXiv:1908.05557 [cs, stat], 1471–1479, https://doi.org/10.1109/ICTAI.2019.00209, 2019.

Wan, Z., Hook, S., and Hulley, G.: MOD11A1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006, 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.

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.

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, Agricultural and Forest Meteorology, 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 Sensing of Environment, 223, 50–62, https://doi.org/10.1016/j.rse.2018.12.029, 2019.

https://doi.org/10.5194/bg-2023-141
Preprint. Discussion started: 31 August 2023

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, 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 Sensing, 10, 327, https://doi.org/10.3390/rs10020327, 2018.

Zhang, Y., Xiao, X., Zhou, S., Ciais, P., McCarthy, H., and Luo, Y.: Canopy and physiological controls of GPP during drought and heat wave, Geophysical Research Letters, 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, https://doi.org/10.48550/arXiv.1904.12054, 26 January 2021.