Contribution of soil algae to the global carbon cycle

Vincent Jassey, Romain Walcker, Paul Kardol, Stefan Geisen, Thierry Heger, Mariusz Lamentowicz, Samuel Hamard, Enrique Lara

Contribution of soil algae to the global carbon cycle

Vincent E.J. Jassey[1]*, Romain Walcker[1], Paul Kardol[2], Stefan Geisen[3,4], Thierry Heger[5], Mariusz Lamentowicz[6], Samuel Hamard[1], Enrique Lara[7]

Affiliations:

[1]Laboratoire Écologie Fonctionnelle et Environnement, Université de Toulouse, CNRS, 31062 Toulouse, France

[2]Department of Forest Ecology and Management, Swedish University of Agricultural Sciences, 90183 Umeå, Sweden

[3]Laboratory of Nematology, Wageningen University, 6708PB Wageningen, The Netherlands

[4]Department of Terrestrial Ecology, Netherlands Institute of Ecology NIOO-KNAW, 6708 PB Wageningen, The Netherlands.

[5]The University of Applied Sciences Western Switzerland, 1260 Changins, Switzerland

[6]Climate Change Ecology Research Unit, Adam Mickiewicz University, 60-001 Poznań, Poland

[7]Real Jardin Botanico, CSIC, Plaza de Murillo 2, 28014 Madrid, Spain

*Correspondence to: Vincent E.J. Jassey; vincent.jassey@univ-tlse3.fr; phone: +33 561 558 923

ORCID IDs

Vincent Jassey, 0000-0002-1450-2437

Romain Walcker, 0000-0002-5769-810X

Paul Kardol, 0000-0001-7065-3435

Stefan Geisen, 0000-0003-0734-727X

Thierry Heger, 0000-0003-3614-0964

Mariusz Lamentowicz, 0000-0003-0429-1530

Samuel Hamard, 0000-0002-9811-4131

Enrique Lara, 0000-0001-8500-522X

Twitter accounts

@vejjassey, @PaulKardol, @stefan_geisen, @utriculator, @HamardSamuel

Word count:

Total: 6767 (Introduction: 811; Method: 3554; Results and discussion: 2402)

4 figures, figures 1, 2 and 3 in color

Supplementary Materials:

Figures S1-S15

Tables S1-S5

Additional references

Summary

- Soil photoautotrophic prokaryotes and micro-eukaryotes– known as soil algae– are, together with heterotrophic microorganisms, a constitutive part of the microbiome in surface soils. Similar to plants, they fix atmospheric carbon (C) through photosynthesis for their own growth, yet their contribution to global and regional biogeochemical C cycling still remains quantitatively elusive.
- Here, we compiled an extensive dataset on soil algae to generate a better understanding of their distribution across biomes and predict their productivity at a global scale by means of machine learning modelling.
- We found that on average $5.5 \pm 3.4 \times 10^6$ algae inhabit each gram of surface soil. Soil algal abundance especially peaked in acidic, moist and vegetated soils. We estimate that globally, soil algae take up around 3.6 Pg C per year, which corresponds to approximately 6% of the net primary production of terrestrial vegetation.
- We demonstrate that the C fixed by soil algae is crucial to the global C cycle and should be integrated into land-based efforts to mitigate C emissions.

Keywords: biogeography; microbial photosynthesis; net primary productivity; photoautotrophs; soil C cycle; soil microbiome

Introduction

Soils are a critical component of the global C cycle and are paramount in mitigating climate change (Amelung *et al.*, 2020). They are the largest repository of organic matter on land, storing ~1,500 Gt C which largely exceeds the amount of C stored in the aboveground vegetation (i.e. ~560 Gt C) (Crowther *et al.*, 2019). The magnitude of the soil organic C pool strongly depends upon microorganisms as microbial growth and activity balance the accumulation and release of organic C through the decomposition of plant litter (Liang *et al.*, 2017a). To date, research on soil microorganisms has mostly focused on heterotrophic microbes and their role in C release with less attention paid to the role of microbial photosynthesis in soil C inputs. Yet, many soil microorganisms are capable of CO_2 fixation (Šantrůčková *et al.*, 2018; Crowther *et al.*, 2019; Oliverio *et al.*, 2020; Akinyede *et al.*, 2020; Bay *et al.*, 2021) and might therefore contribute to soil C fluxes. In particular, while microbial photosynthesis in aquatic systems can quantitatively rival that of terrestrial plants (Field *et al.*, 1998), microbial C fixation in soils has so far never been evaluated at a global scale (but see Elbert *et al.* (2012) for partial estimations based on cryptogam ground cover).

Recent global studies characterizing soil biodiversity have shown that microorganisms capable of CO_2 fixation are omnipresent in soil (Cano-Díaz *et al.*, 2020; Oliverio *et al.*, 2020; Bay *et al.*, 2021). Soil photoautotrophic microbes that fix atmospheric CO_2 through photosynthesis are often referred as soil algae, while others perform CO_2 fixation using chemoautotrophy or heterotrophy via several metabolic pathways and reactions (Miltner *et al.*, 2004). The role of non-phototrophic CO_2 fixation in soil C balance has been increasingly studied in the past few years (Miltner *et al.*, 2004, 2005; Šantrůčková *et al.*, 2018; Spohn *et al.*, 2019; Akinyede *et al.*, 2020). However, soil algae often constitute a small proportion of the soil microbiome biomass (Mitchell *et al.*, 2003; Jassey *et al.*, 2013), and, for this reason, are often seen insignificant for soil C uptake (but see Yuan *et al.*, 2012; Wu *et al.*, 2015; Ge *et al.*, 2016). Yet, soil algae occur in a range of surface soils, such as forest, grassland, and desert soils (Cano-Díaz *et al.*, 2020; Oliverio *et al.*, 2020), and encompass myriads of prokaryotes and micro-eukaryotes, with Cyanobacteria and Chlorophyta being the most commonly reported phyla in soil diversity surveys (Cano-Díaz *et al.*, 2020; Oliverio *et al.*, 2020).

Soil algae have been extensively studied in drylands, where phototrophic biocrusts often constitute the main source of C for the soil system (Maier *et al.*, 2018). However, very few studies have considered the importance of soil C inputs by microscopic algae in other ecosystems (Wyatt *et al.*, 2011; Yuan *et al.*, 2012; Schmidt *et al.*, 2016; Halvorson *et al.*, 2019; Hamard *et al.*, 2021a). Despite their apparent global distribution, our understanding of the ecological preferences of soil algae across broad spatial scales remains limited. While some previous work has suggested that soil moisture availability is a key driver of soil algal net primary productivity (NPP) (Brostoff *et al.*, 2005; Yoshitake *et al.*, 2010; Hamard *et al.*, 2021b,a), other studies have highlighted the importance of

91 temperature (Shimmel & Darley, 1985; Dettweiler-Robinson *et al.*, 2018) or plant community
92 composition (Hamard *et al.*, 2021a), and it remains unclear how predictable soil algal NPP is at larger
93 spatial scales. As a result, quantitative information on soil algal C fixation remain mostly restricted
94 to drylands (Rodriguez-Caballero *et al.*, 2018) and is not readily available at the global scale.
95 Generating quantitative, spatially explicit information about the distribution and productivity of soil
96 algae at a global scale is thus critical for understanding microbial control over soil C dynamics, from
97 contributions to soil C uptake to soil C stabilization and sequestration (Liang *et al.*, 2017a).

98 In this study, we address a set of fundamental questions to advance our understanding of the
99 importance of soil algae in the soil C cycle: 1) What are the habitat preferences of soil algae? 2) How
100 predictable is the CO_2 fixation rate of soil algae across large spatial scales and environmental
101 gradients? 3) What is the contribution of soil algae to ecosystem C uptake? To address these
102 questions, we collated data on soil algal abundance and net primary productivity (NPP) from 203
103 georeferenced locations in all major terrestrial biomes (Fig. 1a; Table S1 and S2). We first conducted
104 a biome-level analysis to reveal the main patterns of soil algal abundance and NPP. Second, we
105 identified the main drivers of the abundance and NPP of soil algae across biomes by using a stack of
106 55 global layers of climate, soil and vegetation characteristics (Table S3). Finally, we used geospatial
107 machine learning to generate a global, one kilometer-resolution map of soil algal net C fixation across
108 the globe, and estimate their global contribution to terrestrial net primary productivity.
109
110 **Materials and Methods**
111 **Literature survey**. We collected data on soil algae from previously published studies and
112 unpublished data collections using a systematic review approach. We searched for studies that
113 quantified the density and/or chlorophyll biomass and/or primary productivity of soil algae in surface
114 soils. Peer-reviewed publications were collected by searching Web of Science (WoS; 1 January 1970
115 to 10 November 2019), Google Scholar (1 January 1970 to 10 November 2019) and ResearchGate to
116 construct our data sets for a period of 7 months (January-June 2019, with updates in November 2019).
117 We used the keywords 'soil algae' OR 'photoautotroph' AND 'biomass' OR 'abundance' AND
118 'photosynthesis' OR 'primary productivity' to build our data base on soil algal abundance, chlorophyll
119 biomass and primary productivity in different type of ecosystems. We standardized our efforts by
120 focusing on studies in which samples were taken from uppermost centimeters soil, including litter, as
121 soil algae further down in the soil column are expected to have only very small photosynthetic rates.
122 For experimental studies, only the controls were considered.
123

Data collection. After initial screening, PDFs of all papers were manually screened to collect data. In order to be suitable for our analyses, the chosen papers had to present (or make reference to) the following information and data:

- Sampled soil algal communities using standard methodologies, which would adequately capture quantitative information of the abundance per gram of dry soil, such as cytometry information, or the C-analyzer used to quantify primary productivity. At a minimum, total abundance or chlorophyll biomass or primary productivity of algae at each site had to be measured. Ideally, there was information on the microbial domain (prokaryotes and/or eukaryotes), with the abundance data (cell counts) of each domain,
- Information on the habitat cover and/or type of ecosystem,
- Available geographic coordinates for all sampled sites, or maps that could be georeferenced. When spatial coordinates were absent, but the type of ecosystem present, we included these data in Fig. 1 but not in further calculations from Fig. 2, 3 and 4.

Data were extracted from tables, figures, the main text and/or supplementary materials; data extraction from figures was performed using Web Plot Digitizer software (https://automeris.io/WebPlotDigitizer/). When multiple values were available in each study, we used the mean in our data analyses. Similarly, when only minimum and maximum values were reported, we used the mean between these two values. Information including publication year, site location, number of plots and ecosystem types were extracted from a total of 166 publications. This resulted in a final subset of 203 georeferenced sites and 19 non-georeferenced sites that were used for further analyses. These sites include a wide range of ecosystem types (forests, grasslands, croplands and wetlands) and climatic regions (arid, temperate, tropical, continental and polar ecosystems), which we classified into seven biomes: grasslands (6.4% of the data), drylands (19.7%), broadleaf and mixed forests (14.3%), alpine and polar lands (20.7%), wetlands (15.3%), croplands (11.8%), and broadleaf evergreen forests (11.8%; see Table S1 and S2 for details).

Data collation. The data taken from one publication, including supplementary material, or from our own unpublished measurements were considered as a 'dataset'. For each dataset, the following site-level community metrics were calculated where possible: total (prokaryotes + micro-eukaryotes) abundance of soil algae per gram of dry soil at the site and soil algal net C uptake in gC m-2 yr-1 at the site level. Issues often arose when compiling data from different studies as the estimate may depend on the methods used. Therefore, we referenced the methodologies used for quantifying soil algal abundance and/or NPP for each dataset.

158 Three methods for quantifying the abundance of soil algae in surface soils were reported in our
159 database: culture-dependent through counting colony-forming units (CFU; 34% of the data; only
160 prokaryotes), flow cytometry (12%), and direct microscopy (54%). These three techniques measure
161 living and active cells only, but their respective limitations (see details in Maron *et al.*, 2006;
162 Kallmeyer *et al.*, 2008; Beal *et al.*, 2020) may have influenced the final estimate of soil algal
163 abundance. Although these techniques give similar trends (Beal *et al.*, 2020), CFU may overestimate
164 cyanobacterial counts as it selects for rapid-growing specimens, whilst flow cytometric and
165 microscopic cell counting may underestimate densities as they require detachment and mechanical or
166 optical separation of cells from interfering soil inorganic particles (Maron *et al.*, 2006). To test the
167 potential bias in our data, we performed a linear mixed effects model with the method used as a fixed
168 effect and microbial domain nested into ecosystem type and latitude as random effect on the intercept.
169 The model did not evidence any influence of the methodology used on abundance data ($P = 0.73$;
170 Fig. S1a).
171 Further, five different methods to estimate soil algal CO_2 fixation were reported in our database:
172 biomass growth quantification (4% of the data), biological oxygen demand (BOD; 4% of the data),
173 chlorophyll fluorescence (4%), infrared CO_2 gas analyzer (82%), and isotopic labelling (^{14}C, 6%).
174 All these methods strictly focused on microbial photosynthetic activity (i.e., CO_2 fixation in the
175 presence of light) and minimized or avoided possible non-phototrophic CO_2 fixation by subtracting
176 dark CO_2 fixation rates from light CO_2 fixation rates. While these techniques differ, they usually are
177 in agreement and give similar trends (Peterson, 1980; Richardson *et al.*, 1984; Hamard *et al.*, 2021a).
178 However, depending of the technique, net photosynthesis of algae can be overestimated (e.g. BOD
179 and chlorophyll fluorescence which provide maximal photosynthetic rates) or underestimated (e.g.
180 ^{14}C isotopic labelling; Richardson1984), especially under conditions of low nutrient and/or high light
181 (Peterson, 1980). Nevertheless, although biomass and isotopic labelling methods tend to overestimate
182 soil algal NPP ($P < 0.05$), we found that this trend was biome-driven and was not significant within
183 biomes ($P = 0.06$; Fig. S1b).
184 Unless mentioned otherwise, soil algal C flux rates from studies providing estimates on an
185 annual basis were included without modification in the data base. Soil algal C flux values reported in
186 μmol, μg, mg, second, minute, hour, or day were converted into g m-2 yr-1. As microbial
187 photosynthesis does not occur every day of the year neither all day long, we constrained the microbial
188 photosynthetic activity to about one third of a day (8 hours), for a limited period of time during the
189 year; i.e. 80 days in arctic areas, 150 days in subarctic areas, 240 days in temperate areas and 300
190 days in tropical areas. Maximal photosynthetic rates under optimal conditions were limited to about
191 1/3 of the maximal value to account for limitations of photosynthetic activity and dark respiration

(Elbert *et al.*, 2012). Reported rates that did not take into account for dark respiration were scaled with a factor of 2/3, as showed by our specific measurements in the field (Fig. S2).

Environmental data collection. In order to identify the main environmental drivers of the abundance and net primary productivity of soil algae, and create spatial predictive models of primary productivity, we sampled a stack of 55 environmental variables at each of the data point locations using the Google Earth Engine Platform (Gorelick *et al.*, 2017) (Table S3). The stack was composed of nineteen long term climate variables extracted from the WorldClim V1 database. They averaged monthly data spanning the period 1960-1991. Fourteen long term additional climate variables were extracted from TerraClimate, and averaged for the period 1958-2019. Several vegetation related variables were extracted from Terra Moderate Resolution Imaging Spectroradiometer (MODIS) : vegetation indices as proxies for plant cover (Leaf Area Index) and biomass (Normalized Difference Vegetation Index, Enhanced Vegetation Index), as well as values of primary productivity (Net Primary Production, Gross Primary Production). We acknowledge that NDVI, EVI, LAI, NPP and GPP are derived from remote sensing reflectance and can thus be sensitive to the chlorophyll fluorescence content of soil algae. However, we assume that under plant coverage, their contribution in surface reflectance remains minor compared to the plant foliage (Chen *et al.*, 2005). Seven soil variables were extracted from the OpenLand database for the 0-5 cm soil depth. Five variables on soil moisture were downloaded from the NASA-USDA Global soil moisture and the NASA-USDA SMAP Global soil moisture. Elevation data were retrieved from the Global Multi-resolution Terrain Elevation Data 2010. Human population density was also integrated in the analysis. Latitudes and longitudes were also integrated in the analyses. All details about environmental variables are given in Table S3.

Data were acquired using the Google Earth Engine platform (Gorelick *et al.*, 2017). Long term statistics (mean, median, minimum and maximum values) were calculated for the whole available period in each database for integration in our numerical analyses. To harmonize the different environmental layers across the globe, it was necessary to aggregate or disaggregate– when appropriate– the spatial resolution of the different layers to match a 30 seconds resolution. Following the spatial harmonization, the global layers were matched with each of the 203 data point locations.

Identification of the variables of importance in driving soil algal abundance and NPP. We used a clustering approach by the *ClustOfVar* package (Chavent *et al.*, 2012) in R to reduce the environmental covariates of interest and select the most representative and least collinear variables. We tested a range of cluster numbers (5, 10, 15 and 20) into the *ClustOfVar* function to define the best number of variables to test in machine learning modelling. Fifteen variables of interest were

identified following *ClustOfVar* (see details in Fig. S3), and related to climate (bio03, bio04, bio10, bio11, bio12, bio14, bio15), vegetation cover (lai), soil (moist, soc, bulk, ph, sand), and geographic (lon) conditions (abbreviations for these variables are given in Table S3 and Fig. 2). We then identified the main environmental drivers of soil algal abundance and NPP among these 15 variables using the Boruta algorithm, a wrapper of random forest and one of the most efficient tools for variable selection (Degenhardt *et al.*, 2019). The Boruta algorithm compares the importance of predictor variables with those of random so-called shadow variables using statistical testing and several runs of random forests. The Boruta algorithm was computed for the entire data set using 1000 iterations.

Predicting soil algal NPP by means of machine learning modelling. Training machine learning (ML) models on a relatively small number of observations can lead to over-fitting and produce inaccurate results. As gathering a bigger dataset to overcome these problems was impossible due to the limited number of available studies in the literature, we used a series of methods and targeted sensitive analyses testing the robustness of our predictions. Particular attention was given to the distribution of data, the number of predictors used, the choice of the model(s) and its (their) hyper-parameters, cross validation strategies, and confidence prediction intervals (Bishop, 2006; Lesmeister, 2019):

1) We inspected the distribution of the soil algal NPP values and searched for outliers as they can strongly influence the model and its prediction. One extremely high and unrealistic value has been removed from the data base (see Table S2 for detail).
2) We implemented our ML models with the most relevant predictors for soil algal NPP based on Boruta analysis (see above). Usually, explicit predictor selection is not the best approach for machine learning, but when data size is limited, it is an essential step to avoid over-fitting (Meyer *et al.*, 2018).
3) Because complex machine learning models with many parameters (e.g., neural networks approaches) are more prone to over-fitting issues (Bishop, 2006; Lesmeister, 2019), we selected and tested six classes of relatively simple machine-learning algorithms (ML) trained on the six best environmental variables identified with Boruta algorithm to spatially predict soil algal NPP across space (see Results). We considered the most common ML models, including simple linear model (GLM), L1-regularization regression linear model (L1-LM), Bayesian Generalized Linear Model (bGLM), k-Nearest-Neighbor (kNN), Bagged MARS (MARS), and Random Forest (RF) as a benchmark. We further combined all ML algorithms into a stack ensemble model (ENS) as predictions from more than one ML algorithm may give better predictive performance than could be obtained from any of the basic or essential learning algorithms alone (Lesmeister, 2019). These ML model types were chosen either

because they were previously used for spatial predictions, or because the general machine-learning literature suggests that they could perform well for this task (Table S4). Each of the ML models included model-specific tuning parameters that were left at default values for initial testing and comparison. To assess predictive performance of the models, we split the total number of points into a training set and a test set using an 80/20 random split. We used the training set to train the different models and the test set to test their performance. We evaluated the model strength using *k*-fold cross-validation (with $k = 10$). For each *k*, we stored the vector of soil algal NPP predictions, which was then used to generate predictive statistics, namely the squared Pearson's correlation between observed soil algal NPP values and those predicted (noted R^2) and the root mean squared error (RMSE). We found that all ML algorithms can successfully predict soil algal NPP, although RF outperformed other ML algorithms by a significant margin (Fig. S4). The ensemble model did not perform better than RF ($P = 0.98$, ANOVA) while giving higher RMSE (Fig. S4). As RF overall performs better and have been demonstrated to provide robust predictions for small sample sizes (Ramezan *et al.*, 2021), we selected RF to create a predictive, high-resolution map of soil algal NPP across the globe as described below.

4) We used a grid-search procedure to iteratively tune the hyperparameters of our RF model in R using *randomForest* and caret packages (Kuhn, 2008): the number of trees to grow ('ntree'; 50, 150, 250, 350 or 450) and the number of variables sampled at each split ('mtry'; 2 to 6), resulting in a total of 25 RF models. The values ntree = 350 and mtry = 4 were defined as the best hyperparameters. We used the training set to determine the best set of model hyperparameters, and to train the model. We used the test-set to assess out-of-sample error, as well as model prediction performance using R^2 and RMSE values as explained above.
5) We used four strategies to cross-validate our best RF model and generate statistics of the model robustness and predictive power (Fig. S5). The first strategy focused on the size of the data set and corresponded to a common *k*-fold cross validation where observations were randomly split into k sets of decreasing size (hereafter, *k*-fold 'size CV') ignoring any structure of potential spatial dependence in the data. Model training was then performed iteratively on k-1 sets. Here, we used $k = 6$; we iteratively and randomly selected 100%, 90%, 80%, 70%, 60% and 50% of the data set to train our random forests models. We choose to maintain the integrity of our data set and not remove a subset of data at the beginning as it would mean the loss of geographic representation. As a test set, we used a data set in which only unique pairs of coordinates were present (65 pairs in total instead of 102 data points; hereafter 'paired dataset'). We summarized our initial dataset using the median applied on similar pairs of coordinates. As local variability could reach 150 $gC.m^{-2}.yr^{-1}$, this approach

297 enabled us to obtain a validation dataset for accuracy assessment. The second strategy, *i.e.*,
298 the *k*-fold 'shuffled CV', is inspired by 'null-model' analyses in ecology, and test the
299 assumption that our predictions are not random and driven by our environmental predictors.
300 To do so, we randomized the environmental predictors matrix to break any structure of
301 environmental dependence in the data. We iteratively and randomly shuffled the
302 environmental matrix 10 times ($k = 10$) before training our random forests model. The third
303 strategy, i.e., *k*-fold 'spatial CV', differs from the size CV in that observations are split into
304 spatially structured clusters (Ploton *et al.*, 2020). Here, the objective was to group
305 observations into spatial clusters and take into consideration the variability generated by the
306 multiple measurements at the same locations, or nearby, in our data set. Spatial clusters were
307 generated using a hierarchical cluster analysis (Ward's hierarchical agglomerative linkage
308 method) of the distance matrix of coordinates and a clustering height of H = 50 km. Here,
309 we used $k = 10$; we iteratively and randomly selected data within each spatial cluster to train
310 our random forests models. The maximum size of the *k*-datasets was 65, *i.e.*, the maximum
311 number of unique pairs of coordinates in our paired dataset. Here again, the paired dataset
312 was used as a test set. The fourth strategy (*k*-fold 'Predictor Variable Shuffling (PVS) CV')
313 tests the assumption that our model gets over-fitted by the covariance among environmental
314 predictors. To refute this assumption, we randomly shuffled the values of 1, 2, 3, 4, and 5
315 predictors before training the RF model. PVS CV was run on spatially clustered training sets
316 (same approach as 'spatial CV') with $k = 100$ to cover all random combinations among
317 predictors. For each CV strategy, we stored the vector of soil algal NPP predictions, which
318 was then used to generate CV statistics, namely the squared Pearson's correlation between
319 observed and predicted NPP values (noted R^2) and the root mean squared error (RMSE).
320 6) To assess any further overfitting and/or highly optimistic evaluations of the predictive power
321 of our RF model due to the spatial dependence in the raw data and model residuals (Ploton
322 *et al.*, 2020), we tested for spatial autocorrelation in the raw data and in size and spatial CV
323 model residuals (Fig. S6). We observed spatial autocorrelation using empirical variograms
324 and did not evidence any particular spatial autocorrelation. The geostatistical analysis *gstat*
325 R package (Pebesma & Heuvelink, 2016) was used for variogram and spatial autocorrelation
326 testing.
327 7) Like many algorithmic approaches to prediction, RF typically produces point predictions
328 that are not accompanied by information about how far those predictions may be from true
329 response values. To cross-validate and quantify this issue, we used prediction intervals that
330 estimate the interval into which future observations will fall with a given probability
331 (Meinshausen, 2006). In other words, it calculates the confidence or certainty in the

prediction. We used 'out-of-the bag' (OOB) prediction intervals as a straightforward approach for constructing our RF prediction intervals (Zhang *et al.*, 2020). As described for the 'spatial CV' strategy, we generated ten independent subsets of our entire data set, stratified by spatial clusters. For each independent subset, we trained our best RF and calculated OOB prediction intervals at each run. Then, we classified whether data points from the test data set fell within or outside RF prediction intervals.

All comparative and cross-validation analyses were performed in R (R Core Team, 2019) using *caret* ensemble packages.

Mapping soil algal NPP and evaluating model uncertainties. To create the final map of soil algal NPP and represent the confidence in our estimates for each pixel, we used an ensemble approach (van den Hoogen *et al.*, 2019; Ma *et al.*, 2021). We averaged the global predictions from ten RF models trained on ten independent subsets of our entire data set, stratified by spatial clusters to proportionally represent the major bioclimatic zones in each of the ten independent subsets ('spatial CV' strategy). This approach minimizes the influence of any single prediction, thereby stabilizing variation and minimizing bias that can otherwise arise from extrapolation or in-fit overfitting when using a single machine learning model (Sagi & Rokach, 2018). The ten independent RF models were run with the six best environmental variables identified through Boruta algorithm and using the best-performing set of hyperparameters (Fig. S7). Through this approach, we thus returned ten times the best RF model using ten different training sets that took into account local variability. We then used the mean predicted value across the ten RF models as the final prediction of soil algal NPP for each pixel. Finally, from these ten models, we further calculated per-pixel coefficient-of-variation values (standard deviation divided by the mean predicted value) as a measure of prediction uncertainty (Ma *et al.*, 2021). In addition, we assessed the extent of extrapolation in our models, that is, how well our sampled data spread throughout the full environmental space, following van den Hoogen *et al.*, (2019). In particular, we examined how many of the Earth's pixels existed outside the range of our sampled data for each of the six environmental layers used in our RF model. To do so, we extracted the minimum and maximum values of each environmental layer of the pixels in which our sampling sites were located. Then, we evaluated for each environmental layer the number of terrestrial pixels that fell outside the sampled range, and calculated the relative proportion of interpolation, that is the percentage of environmental bands that fall into the sampled range. Next, we created a per-pixel representation of the relative proportion of interpolation and extrapolation (Fig. S8). All geospatial and extrapolation analyses were performed in Google Earth Engine (Gorelick *et al.*, 2017).

Cross validation map of soil algal NPP. As an additional validation exercise, we estimated annual soil algal NPP following the biome-based land cover approach taken by Elbert *et al.*, (2012). Soil algal NPP was estimated by multiplying the global ground area surface of a particular biome with its corresponding median of algal C uptake flux (Fig. 1). Biome land covers were recovered from the global land cover characteristics data base version 2.0 available (https://www.usgs.gov/centers/eros/science/usgs-eros-archive-land-cover-products-global-land-cover-characterization-glcc) and reclassified according to our biome classes: grasslands, drylands, broadleaf and mixed forests, alpine and polar lands, wetlands, croplands, and broadleaf evergreen forests.

Results and discussion

Biome-level patterns of soil algal density and NPP. By compiling a dataset on microscopic abundance observations (n = 115; Table S1), we found on average $5.5 \pm 3.4 \times 10^6$ soil algae per gram of dry topsoil (Fig. 1b). Soil algal density varied within and across biomes, ranging from thousands to millions of individuals per gram of dry soil (Fig. 1c). Overall, soil algal abundances (103 cells per gram of dry soil) were highest in wetlands (median = 1036), grasslands (median = 410), broadleaf evergreen forests (median = 202), and croplands (median = 161), while the lowest densities were found in drylands (median = 85), broadleaf and mixed forests (median = 59), and alpine and polar lands (median = 20) (Fig. 1c). These findings show discrepancies with the most recent assessments of the biogeographic distribution of soil algae based on DNA sequencing approaches (Cano-Díaz *et al.*, 2020; Oliverio *et al.*, 2020). These previous studies suggest that soil algae are typically abundant in arid soils, encompassing up to 40% of the total eukaryotic community (Oliverio *et al.*, 2020) and 4% of the total prokaryotic community (Cano-Díaz *et al.*, 2020), respectively. However, even though data based on amplicon gene sequencing give arguably an accurate picture of microbial diversity, our findings illustrate that they cannot be used to infer biogeographic patterns of algal density in soils. Nevertheless, DNA sequencing data could explain the patterns of absolute abundance seen in this study (Hamard *et al.*, 2021a). The blooming of specific taxa resulting from taxonomic turnover in response to specific soil conditions could drastically increase total soil algal abundance (Karaoz *et al.*, 2018).

To identify the main environmental variables that drive algal density in soils, we related the density of soil algae with environmental factors. In contrast to soil invertebrates (van den Hoogen *et al.*, 2019) and total microbial biomass (Xu *et al.*, 2013), our analysis did not reveal notable latitudinal and/or longitudinal effects on soil algal abundance. Instead, and as shown for the community composition of micro-eukaryotes (Oliverio *et al.*, 2020; Aslani *et al.*, 2021) and bacteria (Delgado Baquerizo *et al.*, 2018), we found that climate (i.e., temperature and precipitation) was a main driver

of the global distribution of total algal density in soils. However, plant cover (i.e., leaf area index) and soil characteristics (i.e., soil moisture, soil organic carbon content and pH) were as important as climate (Fig. 2a). In particular, mean annual temperature and precipitation, vegetation cover and soil moisture had strong positive effects, whereas increasing pH had a negative effect on total soil algal abundance (Fig. S9). Although the exact mechanisms behind these interactions still need to be identified, our results indicate that complex interactions among soil properties, climate, and vegetation determine the growth efficiency of soil algae. Our findings suggest that frequent rainfall and plant cover facilitate the size distribution and connectedness of aqueous microbial habitats in soils by increasing water retention in soil pores. This, in turn, promotes microbial cell motility, cell-to-cell interactions (Bickel & Or, 2020), and therefore, algal abundance in top soils. In contrast to the general assumption that arid environments are the main habitat for soil algae (Oliverio *et al.*, 2020), we here show that soil algae are widespread and more abundant in absolute numbers in acidic, wet, and vegetated areas. These results further suggests that the contribution of soil algae to terrestrial productivity is particularly important in these areas.

We tested this assumption by collating a second dataset from the literature on algal NPP in surface soil, and spanning similar biomes as for the abundance data (n = 102; Fig. 1a; Table S2). On average, soil algae were responsible for a mean annual NPP of 30 g C m^{-2} (0.06 - 253.3 g C m^{-2} yr^{-1}) across biomes (Fig. 1b). When comparing these data with terrestrial NPP at the same locations, soil algal NPP accounted for ~10.3% (0.3-80%) of terrestrial NPP (Fig. 1b, c), which is consistent with the value reported by Hamard *et al.*, (2021a) for peatlands (~9.3%). We found the highest algal NPP in croplands (median = 157 gC m^{-2} yr^{-1}), broadleaf and mixed forests (median = 28.2), grasslands (median = 22.6) and broadleaf evergreen forests (median = 18) (Fig. 1c), supporting our empirical and independent observations on absolute abundance (Fig. 1c). We further showed that the absolute abundance and NPP of soil algae were largely driven by the same environmental variables (Fig. 2c), especially soil moisture, vegetation cover and annual precipitation (Fig. S9). The positive correlation between soil algal NPP and increasing vegetation cover might be seen as counterintuitive given that plant canopy reduces the light availability at the soil surface. However, most microscopic algae show optimal photosynthesis at low light intensity (Ritchie & Larkum, 2012; Hamard *et al.,* 2021a) by optimizing light harvesting at low light flux (Perrine *et al.*, 2012). Given the positive relationship between total microbial abundance and metabolic rates in soils (Johnston & Sibly, 2018), our results further presume a simultaneous increase in soil algal abundance and productivity with increasing environmental favorability, which corroborates previous findings in aquatic systems (Liang *et al.*, 2017b).

Global biogeography of soil algal NPP. We implemented an ensemble of ten RF models (Fig. S7) to predict soil algal NPP based on the six best covariate layers (Fig. 2c), *i.e.,* annual precipitation, soil moisture, vegetation cover, soil organic C content, vapor pressure deficit, and soil sand content (see Method). Our ensemble RF models predicted the test data reasonably well (averaged R^2 = 0.51, root-mean-square error RMSE = 0.84, or 2.4 gC m^{-2} yr^{-1}; Fig. S10). It showed a fairly linear relationship with observed soil algal NPP, although predicted NPP tended to be overestimated at low NPP and underestimated at high NPP— a common bias pattern resulting from the RF algorithm (Xu *et al.*, 2016). Nevertheless, a sensitivity analysis based on RF prediction intervals showed that nearly 93% of observations from the test dataset fell within the RF prediction intervals (Fig. S11). This indicates that our RF model provides unbiased results with predictions falling within the full range of observed data. Further rigorous *k*-fold cross-validation steps (Fig. S5) revealed that RF predictions did not lead to over-fitting by the possible covariance among predictors (Fig. S12) and were robust without issues due to the size of the data set (size-CV: R^2 = 0.47 and RMSE = 0.98; Fig. S13 and S14), the spatial structure of the data (spatial-CV: R^2 = 0.39 and RMSE = 1.04; Fig. S13) or possible stochasticity in the predictions (shuffled-CV: R^2 = 0.02 and RMSE = 1.4; Fig. S13). Our cross-validation hence indicates that soil algal NPP can be reasonably predicted while providing accurate predictions within confidence interval and avoiding over-fitting. We therefore used our spatially unbiased RF models ('spatial CV' approach) to upscale observed soil algal NPP across the globe and to map the global distribution of soil algal NPP (Fig. 3).

The quantitative map of soil algal net C fixation showed fixation rates ranging between 2.3 ± 0.3 and 84 ± 32.4 gC m-2 yr-1 (Fig. 3a). Overall, the predictive uncertainty was relatively low, although areas of substantial uncertainty still remain in tropical (central Brazil) and subarctic (north Canada) regions (Fig. 3b). Despite these uncertainties, the map produced through RF modelling provided more detailed and accurate spatial distribution of soil algal NPP than the low-resolution map extrapolated from biome land-cover (Fig. S15). The map did not reveal notable latitudinal trends unlike other soil C processes such as soil respiration (Xu *et al.*, 2013), bacterial and fungal biomass (He *et al.*, 2020) and microbial residence time (He & Xu, 2021). However, it highlighted four hotspots of soil algal NPP (> 50 gC m-2 yr-1) in North-Eastern North America, South-Eastern South America, Central and Western Europe, and Eastern Asia, respectively (Fig. 3a). Further analysis of land cover showed that these four hotspots are dominated by croplands and forests (Fig. 4a), which is in good concordance with our empirical observations (Fig. 1c).

Globally integrated, soil algal NPP amounted to ~3.6 Pg C yr-1 (2.6-4.8 Pg C yr-1; Fig. 4b), which is slightly higher than the value reported for soil cryptogams and that includes bryophytes and lichens (0.34-3.3 Pg C yr-1; Elbert *et al.*, 2012; Porada *et al.*, 2013). Few factors may be responsible for this counterintuitive difference. First, our estimation did not focused on cryptogams ground cover

only (Elbert *et al.*, 2012) but to the whole ground surface. Second, our data compilation on soil algal NPP much exceeds previous efforts and not only included drylands but also many wetter area such as croplands, rainforests and wetlands. These regions evidenced algal NPP values two-to-eight times higher than in drylands (Fig. 1c), on which previous estimations are mostly based (Elbert *et al.*, 2012; Rodriguez-Caballero *et al.*, 2018). Third, previous estimates mostly included cyanobacterial NPP (Elbert *et al.*, 2012). Many cyanobacteria are facultative heterotrophs and often downregulate their own photosynthesis to nearly 10% of their maximum rate when cooccurring with plants as they get carbon from them (Adams & Duggan, 2008).

General implications. With nearly 3.6 Pg C taken up annually, soil algae contribute to ~6.4% of the global terrestrial NPP (~56 Pg C yr^{-1}; Zhao *et al.*, 2005)– again supporting our independent empirical observations across biomes (Fig. 1b). Such contribution to terrestrial NPP might seem relatively high considering the low C biomass of soil algae compared to terrestrial plant biomass (Bar-On *et al.*, 2018). Yet, the photosynthetic capacity of soil algae is comparable to plants when considering their chlorophyll content per unit area (Fig. S16, Table S5). Furthermore, the small fraction of C found in live standing biomass of soil algae does not reflect the amount of C cycled through this pathway as microbial growth rates and turnover are much higher than plants. Accordingly, one might argue that most of the C fixed by soil algae is then rapidly released through respiration and decomposition as soil algae die, minimizing their impact on soil C sequestration. However, recent findings showed that C in soil algal biomass (Yuan *et al.*, 2012) and microbial necromass (Liang *et al.*, 2019) can significantly contribute to soil organic carbon (SOC), thus suggesting that soil algae play a role in soil C sequestration in the long-term. Nevertheless, this contribution to soil C sequestration most probably depends on multifactorial environmental factors (Liang *et al.,* 2019), as well as on the formation and mean residence time of soil algal-derived organic products (Hu *et al.,* 2020). Furthermore, soil algal activity may also initiate hot spots and hot moments of heterotrophic activity in soils by providing resource subsidies to heterotrophic (micro)organisms, either as food for consumers, such as heterotrophic micro-eukaryotes (Seppey *et al.*, 2017), earthworms and springtails (Schmidt *et al.*, 2016) or through the release of labile C that could prime heterotrophic activity (Wyatt & Turetsky, 2015), and hence stimulate decomposition processes in soils (Wyatt & Turetsky, 2015; Halvorson *et al.*, 2019). Although the influence of soil algae in terms of C sequestration and release still remain virtually unknown, our findings indicate that they are important players in global soil carbon uptake, and hence should be taken into account in terrestrial C models.

Despite the confidence in our estimates, we caution that some bias might affect the exact numbers of predicted annual soil algal NPP. First, most algal productivity data rely on snapshot measurements, as continuous, high-resolution measurements of algal NPP in soils are scarce. We thus

used some assumptions to estimate annual soil algal NPP (see Table S2 and Method) and acknowledge that seasonal variation in algal photosynthetic activity due to differences in climatic conditions, light and nutrient availability may influence our estimations. Second, while our model predictions well reflects the observed variation in soil algal NPP across large spatial gradients (e.g., biomes), we acknowledge there are still some limits regarding our ability to accurately predict at smaller spatial scales. Soil algal diversity and community composition (Hamard *et al.*, 2021b,a), predation strength (Schmidt *et al.*, 2016) and/or soil nutrient availability (Gilbert *et al.*, 1998) can influence soil algal activity at small spatial scales. Therefore, some of the unexplained variation in our RF model is probably due to missing plot-level information where soil algal NPP were quantified, explaining the lower range in our RF predictions compared to observations. We minimized this issue by including data from close locations as much as possible. In addition, our spatial-CV taking into account local variability did not show a sharp decline in model R^2, suggesting that local variation does not substantially affect the numbers obtained. Finally, we further note that the size of our data set is limited, as in all other global studies on soil biodiversity (Delgado Baquerizo *et al.*, 2018; van den Hoogen *et al.*, 2019; Cano-Díaz *et al.*, 2020; Oliverio *et al.*, 2020), with some regions being underrepresented. Although our data compilation exceeds previous efforts, and even if we attempted to minimize all issues regarding over-fitting and the size of our data set, the risk that our estimations deviate from the true mean remains, particularly in areas with low sampling density (Fig. 3b). Nevertheless, we tested the extent of extrapolation of our RF model by examining how many of the Earth's pixels existed outside the sampled range of our environmental covariates used in the RF model (see Method). We found that our samples covered the vast majority of environmental conditions on Earth, with 88% of Earth's pixels having at least >80% of the predictor bands falling within the sampled range of environmental conditions (Fig. S8), thus providing confidence in our estimations.

In conclusion, our synthesis presents the most comprehensive assessment of the distribution of soil algae and the global importance of microbial photosynthesis in soil C uptake. Our findings alter some of our most basic assumptions about the role of microorganisms in soil ecological functions by showing that microbial photosynthesis is not only a major component in aquatic ecosystems but also in most terrestrial biomes. We cautiously conclude that soil algae add a so far not considered additional 3.6 Pg C yr^{-1} to net terrestrial C uptake, that is equivalent to roughly 31% of the global anthropogenic C emissions (~11.5 ± 0.9 Pg C; Friedlingstein *et al.*, 2019). Although our estimate of total soil algal NPP will undoubtedly be refined with future data collection, our findings indicate that soil algae are, together with non-phototrophic microbial CO_2 fixation (Spohn *et al.*, 2019; Akinyede *et al.*, 2020), major players in the global soil C balance. Preserving the unseen soil (algal) biodiversity locally and across biomes has never been more important as the urgency to harness all available opportunities to reduce atmospheric CO_2 grows.

Acknowledgments: We acknowledge the work of all researchers that collected these data over the years. This research has been supported by MIXOPEAT, a project funded by the French National Research Agency (grant number ANR-17-CE01-0007) to VEJJ. VEJJ and RW acknowledge financial support from the French National Research Agency through an Investissement d'Avenir (Labex CEBA, ref. ANR-10-LABX-25-01). ML was funded by a grant from the National Science Centre (Poland) (No 2015/17/B/ST10/01656). TH acknowledges support from HES-SO (project 78046, MaLDIveS) and the Swiss Federal Office for the Environment (19.0061.PJ.PZ/D-91173401/988, MiDiBo_2). PK acknowledges financial support from the Swedish Research Council Formas (2017-00366). EL wishes to acknowledge the program "Atracción de Talento Investigador" from the Consejería de Educación, Juventud y Deporte, Comunidad de Madrid, Spain [2017-T1/AMB-5210] and the project MYXOTROPIC funded by the Spanish Government PGC2018-094660-B-I00 (MCIU/AEI/FEDER,UE) for financial support. We thank A. Austin and the three anonymous reviewers for their helpful and constructive comments on our manuscript.

Author contributions: V.E.J.J. developed the study designed with the help of E.L, P.K., M.L. and T.H. V.E.J.J. and T.H. performed the literature review. V.E.J.J collected the data with the contribution of M.L., E.L., T.H., S.G, and S.H. V.E.J.J. and R.W. gathered and organized the data. R.W. collected the environmental and land cover data. V.E.J.J. performed all machine learning analyses. R.W. mapped soil algal productivity with the help of V.E.J.J. S.H. performed additional photosynthesis and chlorophyll biomass analyses on soil algae. V.E.J.J carried out statistical analyses and created the figures. V.E.J.J and E.L. wrote the first draft of the manuscript with inputs from P.K., S.G. and R.W. All authors discussed the results and commented on the manuscript.

Data and materials availability: All environmental co-variates are available online (see Table S3 for details). All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Data and codes related to this paper are available from Figshare (10.6084/m9.figshare.c.5136497).

References

Adams DG, Duggan PS. **2008**. Cyanobacteria–bryophyte symbioses. *Journal of Experimental Botany* **59**: 1047–1058.

Akinyede R, Taubert M, Schrumpf M, Trumbore S, Kuesel K. **2020**. Rates of dark CO2 fixation are driven by microbial biomass in a temperate forest soil. *Soil Biology and Biochemistry* **150**: 1-12.

Amelung W, Bossio D, Vries W, Kögel-Knabner I, Lehmann J, Amundson R, Bol R, Collins C, Lal R, Leifeld J, *et al.* 2020. Towards a global-scale soil climate mitigation strategy. *Nature Communications* **11**: 5427.

Aslani F, Geisen S, Ning D, Tedersoo L, Bahram M. **2021**. Towards revealing the global diversity and community assembly of soil eukaryotes. *Ecology Letters* **25**: 65-76.

Bar-On YM, Phillips R, Milo R. **2018**. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences of the United States of America* **115**: 6506–6511.

Bay SK, Dong X, Bradley JA, Man Leung P, Grinter R, Jirapanjawat T, Arndt SK, M Cook PL, LaRowe DE, Nauer PA, *et al.* 2021. Trace gas oxidizers are widespread and active members of soil microbial communities. *Nature Microbiology* **6**: 246–256.

Beal J, Farny NG, Haddock-Angelli T, Selvarajah V, Baldwin GS, Buckley-Taylor R, Gershater M, Kiga D, Marken J, Sanchania V, *et al.* 2020. Robust estimation of bacterial cell count from optical density. *Communications biology* **3:512**: 1–29.

Bickel S, Or D. **2020**. Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes. *Nature Communications* **11**: 116–119.

Bishop CM. **2006**. *Pattern recognition and machine learning*. Information Science and Statistics Series. Springer New York, 758 p.

Brostoff WN, Rasoul Sharifi M, Rundel PW. **2005**. Photosynthesis of cryptobiotic soil crusts in a seasonally inundated system of pans and dunes in the western Mojave Desert, CA: Field studies. *Flora - Morphology, Distribution, Functional Ecology of Plants* **200**: 592–600.

Cano-Díaz C, Maestre FT, Eldridge DJ, Singh BK, Bardgett RD, Fierer N, Delgado Baquerizo M. **2020**. Contrasting environmental preferences of photosynthetic and non-photosynthetic soil cyanobacteria across the globe . *Global Ecology and Biogeography* **29**: 2025-2038.

Chavent M, Simonet VK, Liquet B, Saracco J. **2012**. ClustOfVar: An R Package for the Clustering of Variables. *Journal of statistical software* **50**: 1–16.

Chen J, Zhang MY, Wang L, Shimazaki H, Tamura M. **2005**. A new index for mapping lichen-dominated biological soil crusts in desert areas. *Remote Sensing of Environment* **96**: 165–175.

Crowther TW, van den Hoogen J, Wan J, Mayes MA, Keiser AD, Mo L, Averill C, Maynard DS. **2019**. The global soil community and its influence on biogeochemistry. *Science* **365**: eaav0550.

Degenhardt F, Seifert S, Szymczak S. **2019**. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in bioinformatics* **20**: 492–503.

Delgado Baquerizo M, Oliverio AM, Brewer TE, Benavent-Gonzalez A, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK, Fierer N. **2018**. A global atlas of the dominant bacteria found in soil. *Science* **359**: 320–325.

Dettweiler-Robinson E, Nuanez M, Litvak ME. **2018**. Biocrust contribution to ecosystem carbon fluxes varies along an elevational gradient. *Ecosphere* **9**: e02315-12.

Elbert W, Weber B, Burrows S, Steinkamp J, Büdel B, Andreae MO, Pöschl U. **2012**. Contribution of cryptogamic covers to the global cycles of carbon and nitrogen. *Nature geoscience* **5**: 459–462.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. **1998**. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**: 237–240.

Friedlingstein P, Jones MW, O'Sullivan M, Andrew RM, Hauck J, Peters GP, Peters W, Pongratz J, Sitch S, Le Quéré C, *et al.* 2019. Global carbon budget 2019. *Earth System Science Data* **11**: 1783–1838.

Ge T, Wu X, Liu Q, Zhu Z, Yuan H, Wang W, Whiteley AS, Wu J. **2016**. Effect of simulated tillage on microbial autotrophic CO2 fixation in paddy and upland soils. *Scientific Reports* **6**: 1–9.

Gilbert D, Amblard C, Bourdier G, Francez A. **1998**. The Microbial Loop at the Surface of a Peatland:Structure, Function, and Impact of Nutrient Input. **35**: 83–93.

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. **2017**. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* **202**: 18–27.

Halvorson HM, Barry JR, Lodato MB, Findlay RH, Francoeur SN, Kuehn KA. **2019**. Periphytic algae decouple fungal activity from leaf litter decomposition via negative priming (DC Allen, Ed.). *Functional Ecology* **33**: 188–201.

Hamard S, Céréghino R, Barret M, Sytiuk A, Lara E, Dorrepaal E, Kardol P, Küttim M, Lamentowicz M, Leflaive J, *et al.* 2021a. Contribution of microbial photosynthesis to peatland carbon uptake along a latitudinal gradient. *Journal of Ecology* **109**: 1365-2745.

Hamard S, Küttim M, Céréghino R, Jassey VEJ. **2021b**. Peatland microhabitat heterogeneity drives phototrophic microbes distribution and photosynthetic activity. *Environmental Microbiology* **23**: 6811-6827.

He L, Mazza Rodrigues JL, Soudzilovskaia NA, Barceló M, Olsson PA, Song C, Tedersoo L, Yuan F, Yuan F, Lipson DA, *et al.* 2020. Global biogeography of fungal and bacterial biomass carbon in topsoil. *Soil Biology and Biochemistry* **151**: 108024.

He L, Xu X. **2021**. Mapping soil microbial residence time at the global scale. *Global Change Biology*

27: 6484-6497.

van den Hoogen J, Geisen S, Routh D, Ferris H, Traunspurger W, Wardle DA, de Goede RGM, Adams BJ, Ahmad W, Andriuzzi WS, *et al.* 2019. Soil nematode abundance and functional group composition at a global scale. *Nature* **2**: 1042.

Hu Y, Zheng Q, Noll L, Zhang S, Wanek W. **2020**. Direct measurement of the in situ decomposition of microbial-derived soil organic matter. *Soil Biology and Biochemistry* **141**: 107660.

Jassey VEJ, Chiapusio G, Binet P, Buttler A, Laggoun-Défarge F, Delarue F, Bernard N, Mitchell EAD, Toussaint M-L, Francez A-J, *et al.* 2013. Above- and belowground linkages in Sphagnum peatland: climate warming affects plant-microbial interactions. *Global Change Biology* **19**: 811–823.

Johnston ASA, Sibly RM. **2018**. The influence of soil communities on the temperature sensitivity of soil respiration. *Nature Ecology & Evolution* **2**: 1597–1602.

Kallmeyer J, Smith DC, Spivack AJ, D'Hondt S. **2008**. New cell extraction procedure applied to deep subsurface sediments. *Limnology and Oceanography: Methods* **6**: 236–245.

Karaoz U, Couradeau E, da Rocha UN, Lim H-C, Northen T, Garcia-Pichel F, Brodie EL, Bailey MJ. **2018**. Large Blooms of Bacillales (Firmicutes) Underlie the Response to Wetting of Cyanobacterial Biocrusts at Various Stages of Maturity (MJ Bailey, Ed.). *mBio* **9**: e01366-16.

Kuhn M. **2008**. Building predictive models in R using the caret package. *Journal of Statistical Software* **28**: 1–26.

Lesmeister C. **2019**. *Mastering machine learning with R : advanced machine learning techniques for building smart applications with R 3.5, 3rd edition*. Packt Publishing Limited, United Kingdom, 356 p.

Liang C, Amelung W, Lehmann J, Kaestner M. **2019**. Quantitative assessment of microbial necromass contribution to soil organic matter. *Global Change Biology* **25**: 3578–3590.

Liang C, Schimel JP, Jastrow JD. **2017a**. The importance of anabolism in microbial control over soil carbon storage. *Nature Microbiology* **2**: 17105–17106.

Liang Y, Zhang Y, Wang N, Luo T, Zhang Y, Rivkin RB. **2017b**. Estimating Primary Production of Picophytoplankton Using the Carbon-Based Ocean Productivity Model: A Preliminary Study. *Frontiers in microbiology* **8**: 12–75.

Ma H, Mo L, Crowther TW, Maynard DS, van den Hoogen J, Stocker BD, Terrer C, Zohner CM. **2021**. The global distribution and environmental drivers of aboveground versus belowground plant biomass. *Nature Ecology and Evolution* **5**: 1110–1122.

Maier S, Tamm A, Wu D, Caesar J, Grube M, Weber B. **2018**. Photoautotrophic organisms control microbial abundance, diversity, and physiology in different types of biological soil crusts. *The ISME Journal* **12**: 1032–1046.

Maron PA, Schimann H, Ranjard L, Brothier E, Domenach AM, Lensi R, Nazaret S. **2006**. Evaluation of quantitative and qualitative recovery of bacterial communities from different soil types by density gradient centrifugation. *European Journal of Soil Biology* **42**: 65–73.

Meinshausen N. **2006**. Quantile Regression for Left-Truncated Semicompeting Risks Data. *Journal of Machine Learning Research* **7**: 983–999.

Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T. **2018**. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software* **101**: 1–9.

Miltner A, Kopinke F-D, Kindler R, Selesi D, Hartmann A, Kästner M. **2005**. Non-phototrophic CO_2 fixation by soil microorganisms. *Plant and Soil* **269**: 193–203.

Miltner A, Richnow H-H, Kopinke F-D, Kästner M. **2004**. Assimilation of CO2 by soil microorganisms and transformation into soil organic matter. *Organic Geochemistry* **35**: 1015–1024.

Mitchell EAD, Gilbert D, Buttler A, Amblard C, Grosvernier P, Gobat JM. **2003**. Structure of microbial communities in Sphagnum peatlands and effect of atmospheric carbon dioxide enrichment. **46**: 187–199.

Oliverio AM, Geisen S, Delgado Baquerizo M, Maestre FT, Turner BL, Fierer N. **2020**. The global-scale distributions of soil protists and their contributions to belowground systems. *Science advances* **6**: eaax8787.

Pebesma E, Heuvelink G. **2016**. Spatio-temporal interpolation using gstat. *The R Journal* **8**: 204–218.

Perrine Z, Negi S, Sayre RT. **2012**. Optimization of photosynthetic light energy utilization by microalgae. *Algal Research* **1**: 134–142.

Peterson BJ. **1980**. Aquatic Primary Productivity and the 14C-CO2 Method: A History of the Productivity Problem. *Annual Review of Ecology and Systematics* **11**: 359–385.

Ploton P, Mortier F, Rejou-Mechain M, Barbier N, Picard N, Rossi V, Dormann C, Cornu G, Viennois G, Bayol N, *et al.* 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications* **11**: 4511–4540.

Porada P, Weber B, Elbert W, Poeschl U, Kleidon A. **2013**. Estimating global carbon uptake by lichens and bryophytes with a process-based model. *Biogeosciences* **10**: 6989–7033.

R Core Team. **2019**. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramezan CA, Warner TA, Maxwell AE, Price BS. **2021**. Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data. *Remote Sensing 2021, Vol. 13, Page 368* **13**: 368.

Richardson K, Samuelsson G, Hällgren JE. **1984**. The relationship between photosynthesis

measured by 14C incorporation and by uptake of inorganic carbon in unicellular algae. *Journal of Experimental Marine Biology and Ecology* **81**: 241–250.

Ritchie RJ, Larkum AWD. **2012**. Modelling photosynthesis in shallow algal production ponds. *Photosynthetica* **50**: 481–500.

Rodriguez-Caballero E, Belnap J, Büdel B, Crutzen PJ, Andreae MO, Pöschl U, Weber B. **2018**. Dryland photoautotrophic soil surface communities endangered by global change. *Nature geoscience* **11**: 185-189.

Sagi O, Rokach L. **2018**. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**: 1–18.

Šantrůčková H, Kotas P, Bárta J, Urich T, Čapek P, Palmtag J, Eloy Alves RJ, Biasi C, Diáková K, Gentsch N, *et al.* 2018. Significance of dark CO2 fixation in arctic soils. *Soil Biology and Biochemistry* **119**: 11–21.

Schmidt O, Dyckmans J, Schrader S. **2016**. Photoautotrophic microorganisms as a carbon source for temperate soil invertebrates. *Biology letters* **12**: 20150646.

Seppey CVW, Singer D, Dumack K, Fournier B, Belbahri LL, Mitchell EAD, Lara E. **2017**. Distribution patterns of soil microbial eukaryotes suggests widespread algivory by phagotrophic protists as an alternative pathway for nutrient cycling. *Soil Biology and Biochemistry* **112**: 68–76.

Shimmel SM, Darley WM. **1985**. Productivity and Density of Soil Algae in an Agricultural System. *Ecology* **66**: 1439–1447.

Spohn M, Mueller K, Hoeschen C, Mueller CW, Marhan S. **2019**. Dark microbial CO2 fixation in temperate forest soils increases with CO2 concentration. *Global Change Biology* **26**: 1935.

Wu X, Ge T, Wang W, Yuan H, Wegner C-E, Zhu Z, Whiteley AS, Wu J. **2015**. Cropping systems modulate the rate and magnitude of soil microbial autotrophic CO2 fixation in soil. *Frontiers in microbiology* **6**: 379.

Wyatt KH, Turetsky MR. **2015**. Algae alleviate carbon limitation of heterotrophic bacteria in a boreal peatland (R Aerts, Ed.). *Journal of Ecology* **103**: 1165–1171.

Wyatt KH, Turetsky MR, Rober AR, Giroldo D, Kane ES, Stevenson RJ. **2011**. Contributions of algae to GPP and DOC production in an Alaskan fen: effects of historical water table manipulations on ecosystem responses to a natural flood. *Oecologia* **169**: 821–832.

Xu L, Saatchi SS, Yang Y, Yu Y, White L. **2016**. Performance of non-parametric algorithms for spatial mapping of tropical forest structure. *Carbon balance and management* **11**: 14–18.

Xu X, Thornton PE, Post WM. **2013**. A global analysis of soil microbial biomass carbon, nitrogen and phosphorus in terrestrial ecosystems. *Global Ecology and Biogeography* **22**: 737–749.

Yoshitake S, Uchida M, Koizumi H, Kanda H, Nakatsubo T. **2010**. Production of biological soil crusts in the early stage of primary succession on a High Arctic glacier foreland. *The New phytologist* **186**: 451–460.

Yuan H, Ge T, Chen C, O'Donnell AG, Wu J. **2012**. Significant Role for Microbial Autotrophy in the Sequestration of Soil Carbon. *Applied and environmental microbiology* **78**: 2328–2336.

Zhang H, Zimmerman J, Nettleton D, Nordman DJ. **2020**. Random Forest Prediction Intervals. *American Statistician* **74**: 392–406.

Zhao M, Heinsch FA, Nemani RR, Running SW. **2005**. Improvements of the MODIS terrestrial gross and net primary production global data set. *Remote Sensing of Environment* **95**: 164–176.

Supplementary materials

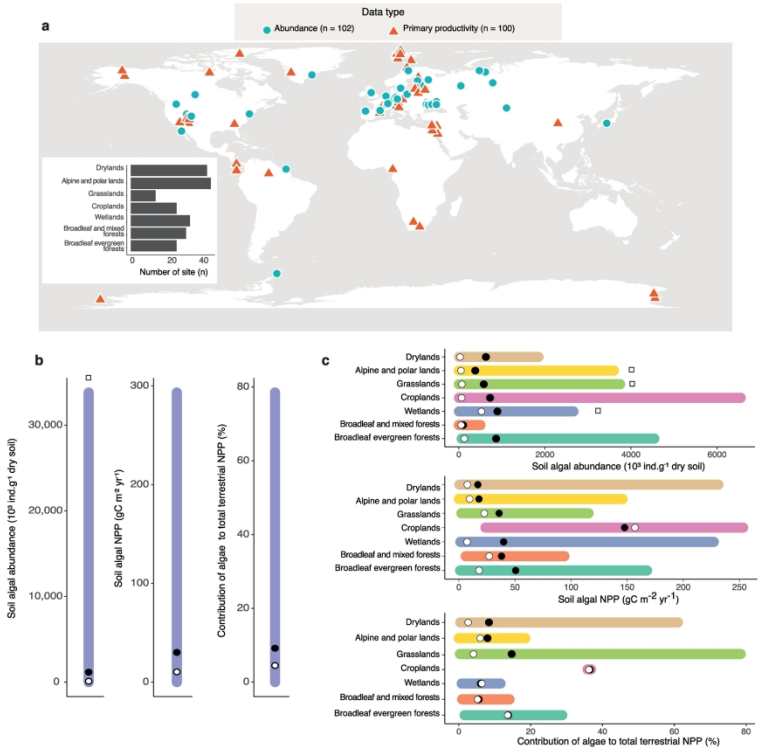
Figure captions

Fig. 1. Sample locations and abundance, annual net primary productivity (NPP) and contribution of soil algae to total NPP. (**a)**, A total of 203 georeferenced data points were collected from the literature and unpublished data, and grouped into biome categories. (**b)**, Averaged soil algal abundance (n = 101), NPP (n = 102), and contribution to total NPP (n = 100), overall, (**c)**, and per biome categories. Open circles represent the mean and filled circles the median. Coloured bars indicate the range between the minimum and maximum values. Little squares indicate clipped extreme values; 1 clipped value in b) and 3 clipped values in c). These extremely high values have been removed to increase the readability of means and medians.

Fig. 2. Environmental factors controlling the abundance and net primary productivity (NPP) of soil algae. **(a, b)**, Results from Boruta algorithm evaluating the relevance of different environmental predictors for the abundance and NPP of soil algae. Arrow lengths represent the mean relevance of each predictor variable, whereas shaded areas represent the maximum importance of each variable. The matrix of environmental predictors was reduced beforehand to select the most representative and least collinear variables (Fig. S3, see Materials and Methods section).

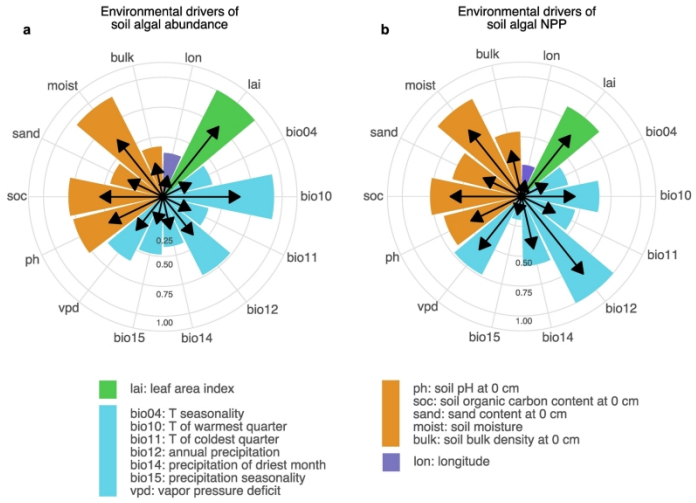
Fig. 3: Global map of soil algal net primary productivity at the 30 arcsec pixel scale (approximately 1 km^2). (**a**) Net primary productivity (NPP) of soil algae in surface soil (gC per square meter per year). Pixel values were binned into fifteen quantiles to create the color palette. Grey color indicates non-investigated area. **(b)** Coefficient of variation (standard deviation as a fraction of the mean predicted value) as a measure of soil algal primary productivity prediction accuracy. Overall, our prediction error is the low at the exception of low soil algal C flux rate in tropical forests and boreal zones. Pixel values were binned into fifteen quantiles to create the color palette.

Fig. 4: Dominant land uses in soil algal NPP hotspots and total soil algal NPP. (**a**) The mean and standard error (n = 10) of the land uses identified in the four hotspots of soil algal NPP (>50 g.m^{-2}.yr^{-1}) from the 10 prediction maps (Fig. S7) used to build the final map presented in Fig. 3a. (**b**) The median and interquantile range (n = 10) of the total soil algal productivity per year from the 10 ensemble random forest models used in the prediction maps presented in Fig. 3a.



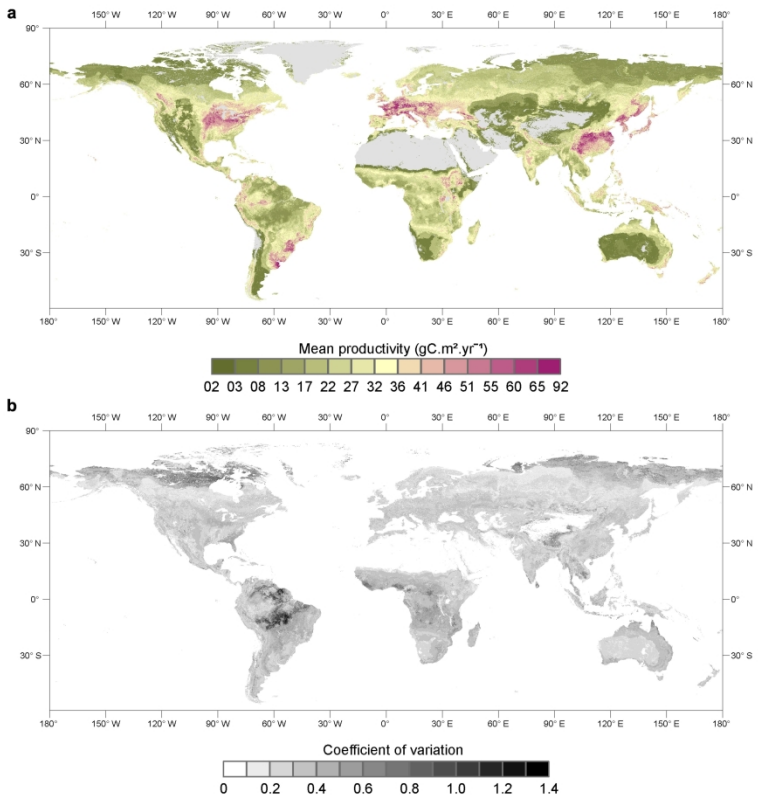
Sample locations and abundance, annual net primary productivity (NPP) and contribution of soil algae to total NPP. (a), A total of 203 georeferenced data points were collected from the literature and unpublished data, and grouped into biome categories. (b), Averaged soil algal abundance (n = 101), NPP (n = 102), and contribution to total NPP (n = 100), overall, (c), and per biome categories. Open circles represent the mean and filled circles the median. Coloured bars indicate the range between the minimum and maximum values. Little squares indicate clipped extreme values; 1 clipped value in b) and 3 clipped values in c). These extremely high values have been removed to increase the readability of means and medians.

209x296mm (300 x 300 DPI)



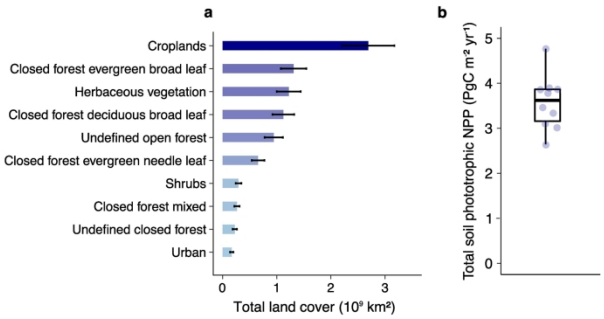
Environmental factors controlling the abundance and net primary productivity (NPP) of soil algae. (a, b), Results from Boruta algorithm evaluating the relevance of different environmental predictors for the abundance and NPP of soil algae. Arrow lengths represent the mean relevance of each predictor variable, whereas shaded areas represent the maximum importance of each variable. The matrix of environmental predictors was reduced beforehand to select the most representative and least collinear variables (Fig. S3, see Method).

209x296mm (300 x 300 DPI)



Global map of soil algal net primary productivity at the 30 arcsec pixel scale (approximately 1 km2). (a) Net primary productivity (NPP) of soil algae in surface soil (gC per square meter per year). Pixel values were binned into fifteen quantiles to create the color palette. Grey color indicates non-investigated area. (b) Coefficient of variation (standard deviation as a fraction of the mean predicted value) as a measure of soil algal primary productivity prediction accuracy. Overall, our prediction error is the low at the exception of low soil algal C flux rate in tropical forests and boreal zones. Pixel values were binned into fifteen quantiles to create the color palette.

210x297mm (300 x 300 DPI)



Dominant land uses in soil algal NPP hotspots and total soil algal NPP. (a) The mean and standard error (n = 10) of the land uses identified in the four hotspots of soil algal NPP (>50 g.m-2.yr-1) from the 10 prediction maps (Fig. S7) used to build the final map presented in Fig. 3a. (b) The median and interquantile range (n = 10) of the total soil algal productivity per year from the 10 ensemble random forest models used in the prediction maps presented in Fig. 3a.

209x296mm (300 x 300 DPI)