Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-023-26824-6

RESEARCH ARTICLE

Machine learning for prediction of soil CO_2 emission in tropical forests in the Brazilian Cerrado

Kleve Freddy Ferreira Canteral[1] · Maria Elisa Vicentini[1] · Wanderson Benerval de Lucena[1] · Mário Luiz Teixeira de Moraes[2] · Rafael Montanari[1] · Antonio Sergio Ferraudo[1] · Nelson José Peruzzi[1] · Newton La Scala Jr.[1] · Alan Rodrigo Panosso[1]

Received: 29 November 2022 / Accepted: 1 April 2023

Abstract

Soil CO_2 emission (FCO_2) is a critical component of the global carbon cycle, but it is a source of great uncertainty due to the great spatial and temporal variability. Modeling of soil respiration can strongly contribute to reducing the uncertainties associated with the sources and sinks of carbon in the soil. In this study, we compared five machine learning (ML) models to predict the spatiotemporal variability of FCO_2 in three reforested areas: eucalyptus (RE), pine (RP) and native species (RNS). The study also included a generalized scenario (GS) where all the data from RE, RP and RNS were included in one dataset. The ML models include generalized regression neural network (GRNN), radial basis function neural network (RBFNN), multilayer perceptron neural network (MLPNN), adaptive neuro-fuzzy inference system (ANFIS) and random forest (RF). Initially, we had 32 attributes and after pre-processing, including Pearson's correlation, canonical correlation analysis (CCA), and biophysical justification, only 21 variables remained. We used as input variables 19 soil properties and climate variables in reforested areas of eucalyptus, pine and native species. RF was the best model to predict soil respiration to RE [adjusted coefficient of determination (R^2 adj): 0.70 and root mean square error (RMSE): 1.02 μmol m^{-2} s^{-1}], RP (R^2 adj: 0.48 and RMSE: 1.07 μmol m^{-2} s^{-1}) and GS (R^2 adj: 0.70 and RMSE: 1.05 μmol m^{-2} s^{-1}). Our findings support that RF and GRNN are promising for predicting soil respiration of reforested areas which could help to identify and monitor potential sources and sinks of the main additional greenhouse gas over ecosystems.

Keywords Soil respiration · Climate change · Environmental modeling · Tropical ecosystems

Responsible Editor: V.V.S.S. Sarma

Highlights

- Tropical forest types were highly sensitive to input variables (climate and soil).
- Random forest is suitable to predict soil CO_2 emissions (FCO_2) from tropical forests.
- Adaptive neuro-fuzzy inference system is not recommended for FCO_2 prediction.

✉ Kleve Freddy Ferreira Canteral
canteralkleve@gmail.com

[1] Department Engineering and Exact Sciences, School of Agricultural and Veterinarian Sciences, São Paulo State University (FCAV/UNESP), Via de Acesso Prof. Paulo Donato Castellane S/N, Jaboticabal, São Paulo 14884-900, Brazil

[2] Department of Phytotecnics, Faculty of Engineer (FEIS/UNESP), Avenida Brasil – Centro, Ilha Solteira, São Paulo 15385-000, Brazil

Introduction

Anthropogenic greenhouse gas (GHG) emissions have increased considerably due to the increasing use of fossil fuels, deforestation and land use and cover change (LUCC) (UNFCCC 2013). Carbon dioxide (CO_2) is one of the main gases that contributes to the additional greenhouse effect and it is global mean concentration increased approximately 47% above the levels in 1750 (representative of the pre-industrial era) (IPCC 2021). It is estimated that between 1990 and 2010, agriculture, forestry and LUCC were responsible for ~ 10–12 Gt CO_2 eq yr^{-1} of global anthropogenic GHG emissions, which resulted primarily from deforestation and agricultural practices (Smith et al. 2014; Tubiello et al. 2014). According to SEEG (2022), Brazil emitted a total of 71 Gt CO_2 eq $year^{-1}$ between 1990 and 2021, where 40.5 Gt CO_2 eq $year^{-1}$ (57%) are linked

Published online: 12 April 2023

to LUCC. In this context, in 2010, Brazil implemented the Plan of Low Carbon Agriculture (LCA) with the aim of promoting sustainable production technologies in line with the country's commitments to reduce GHG emissions in the agricultural sector.

The Brazilian Cerrado biome, also known as the Brazilian savanna, is the second largest ecosystem in South America. This ecosystem encompas an area of approximately 2 million km^2 in the center of the country and is globally important due to its biodiversity and high endemism rates (Anache et al. 2018). However, in the last 50 years, more than half of the native forests in this biome have been converted into pastures and agricultural areas (Vicentini et al. 2019). Deforestation of native forests causes imbalances in ecological relationships and ecosystem processes, which affect CO_2 flows between environmental compartments and drive total GHG emissions in the country (Cerri et al. 2018). For this reason, management practices that maintain or increase soil organic matter (SOM) content are essential to reduce soil carbon losses and at the same time mitigate GHG emissions (Ussiri and Lal 2009).

Soil respiration or soil CO_2 emission (FCO_2) is a critical component of the global carbon cycle, but it is a source of great uncertainty due to the great spatial and temporal variability (Grunwald 2022). FCO_2 is a biochemical process that results from the mineralization of soil organic matter from the biological activity of microorganisms (Ussiri and Lal 2009). Generally, soil temperature and moisture are the main drivers of the spatial and temporal variability of this phenomenon (Tavares et al. 2016; Freitas et al. 2018; Moitinho et al. 2021; Zhang et al. 2022). However, chemical and physical properties of soil are closely linked to the production and transport of gases from the soil to the atmosphere, respectively (Farhate et al. 2018; Silva et al. 2019). In addition, meteorological conditions such as precipitation, relative humidity and air temperature can also affect the process of carbon loss from the soil to the atmosphere (Hamrani et al. 2020; Abbasi et al. 2021). Soil CO_2 emissions are generally measured experimentally using closed chambers or conventional statistical approaches. However, long-term data collection and field monitoring of GHG is very labor-intensive and expensive and conventional statistical techniques follow probabilistic assumptions, in which the results may be associated with uncertainties of various magnitudes (Khan and Khan 2019). To circumvent these limitations, interest in using machine learning (ML) models has significantly increased to simulate complex phenomena (Zendehboudi et al. 2018).

Machine learning (ML) approach are increasingly used to study complex environmental phenomena with high variability in space and time (Zendehboudi et al. 2018; Hamrani et al. 2020). Soil carbon modeling using ML allows for explicit and rigorous assessments using various error metrics and assessing uncertainties associated with the global carbon cycle (Grunwald 2022). In recent years, these techniques have demonstrated satisfactory performance in modeling complex problems and fitting models to continuous nonlinear functions in the area of soil sciences (Abbasi et al. 2021; Kumari and Singh 2022; Singh et al. 2022). Such models include shallow neural networks, such as general regression neural network (GRNN) (Chen et al. 2018; Zhang et al. 2022), radial basis function neural network (RBFNN) (Kashi et al. 2014; Najafi et al. 2018) and multilayer perceptron neural network (MLPNN) (Freitas et al. 2018; Fernandes et al. 2020). In addition, classical regression models such as random forest (RF) (Hamrani et al. 2020; Abbasi et al. 2021; Kumari and Singh 2022; Singh et al. 2022) and hybrid models such as adaptive neuro-fuzzy inference system (ANFIS) (Gopalakrishnan et al. 2011; Najafi et al. 2018; Khan and Khan 2019) have been used to model complex interactions between independent variables.

Previous efforts have been applied to quantify and model variables of environmental interest using ML algorithms (Philibert et al. 2013; Kashi et al. 2014; Freitas et al. 2018; Zendehboudi et al. 2018). Nevertheless, there is a lack of research on the use ML models the spatial and temporal variability of soil CO_2 emissions in tropical forests. As such, the hypothesis of this study is that land use can alter the physicochemical and biological attributes of soil and consequently the spatiotemporal patterns of soil CO_2 emission. Therefore, we sought to evaluate the performance of ML regression models to predict the spatiotemporal variability of FCO_2 in three reforested areas: eucalyptus (RE), pine (RP) and native species (RNS). The study also included a generalized scenario (GS) where all the data from RE, RP and RNS were included in one dataset. The comparison includes three shallow neural network models: GRNN, RBFNN e MLPNN; one classical regression model: RF and one hybrid model: ANFIS. In addition, this study also identified the best chemical and physical properties of the soil and climate attributes for soil CO_2 emission prediction.

Material and methods

Field sampling

The dataset is the result of a field study conducted between November 2015 and May 2016, in forest areas planted with eucalyptus (*Eucalyptus camaldulensis*) and pine (*Pine caribaea* var. *hondurensis*) trees in 1986, and reforested with 35 native species randomly distributed in the area. These areas are located on the Experimental Farm of the UNESP School of Engineering, in the city of Selvíria, Mato Grosso do Sul state, Brazil, on the banks of the Parana River (Fig. 1) (20º 20′ 53.41″ South and 51° 23′ 55.50’’ West, 354 m above sea

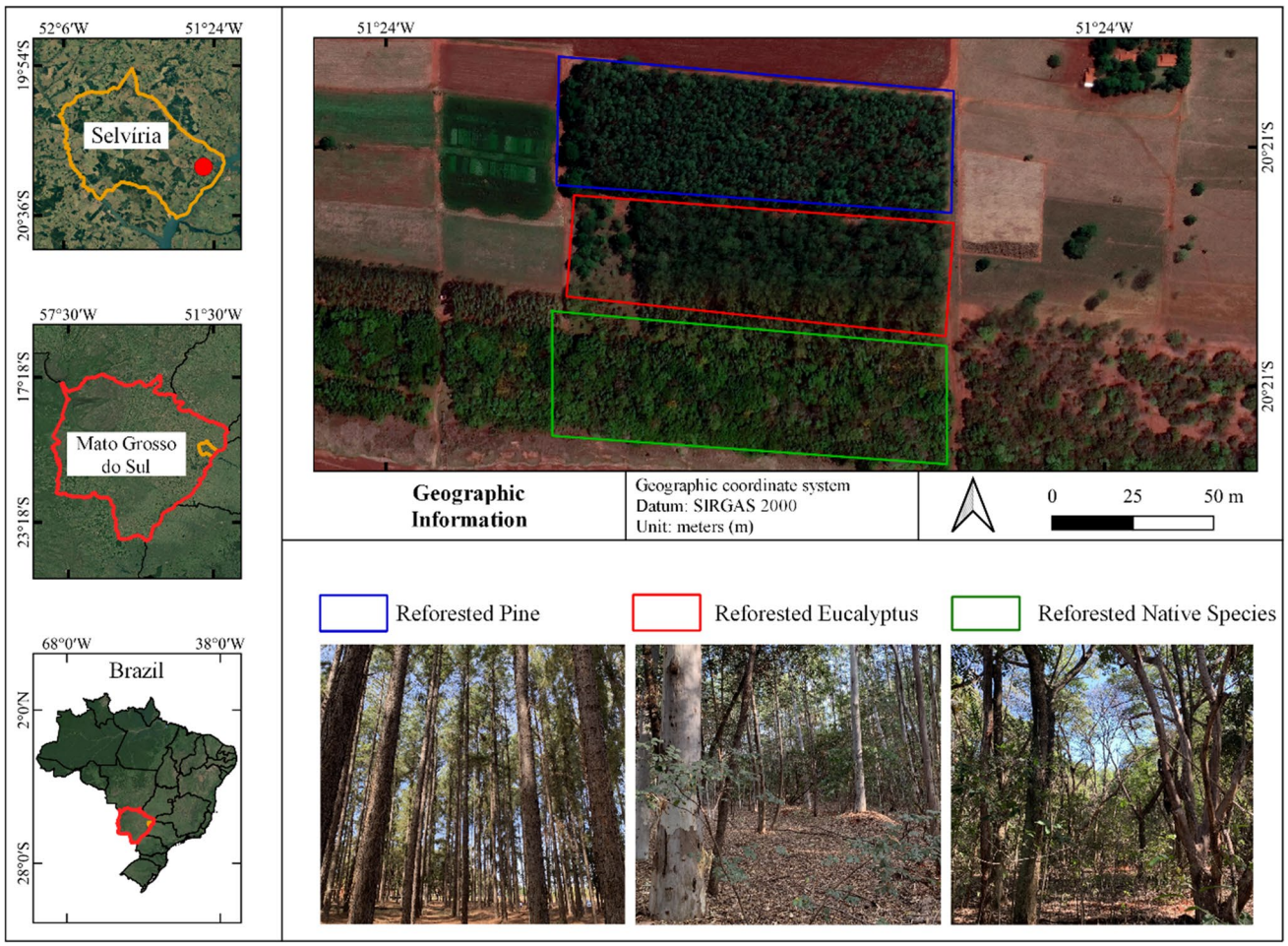

Fig. 1 Map of the study area located in the municipality of Selvíria, Mato Grosso do Sul (MS), Brazil

level). The soil was classified as dystrophic Oxisol (Haplic Acrustox). The topography of the region is characterized as moderately flat and undulating, originating from basaltic soils, according to Soil Taxonomy (Soil Survey Staff 2014). The climate is tropical wet (Aw) according to Köppen's classification system, with a wet summer and dry winter, average annual rainfall of 1370 mm, temperature of 23.5 ºC and relative humidity between 70 and 80%.

Determination of soil CO_2 emission, temperature and soil moisture

The measurements for each experimental plot were made on 20 days over the 193-day study period between November 2015 and May 2016. Assessments of the study areas were performed on the same day in the morning (7am-12 pm). Soil CO_2 emissions were recorded using a soil flux system (LI-8100; LI-COR Bioscience, Nebraska, USA). A total of 25 sampling points were established in each area, using polyvinyl chloride (PVC) collars with a diameter of 0.10 m and height of 0.085 m; the collars remained fixed throughout the experiment. Soil temperature was determined with a digital thermometer and soil moisture (Ms) was measured by a Time Domain Reflectometry (TDR) system (Hydrosense TM, Campbell Scientific Inc., Logan, UT, USA), which consists of two 0.12 m probes that are inserted into the soil, also at 0.10 m from the PVC collars.

Determination of soil chemical and physical variables

Soil samples were collected at a depth of 0.0 to 0.10 m in order to determine soil physicochemical properties. Chemical (phosphorous (P), potassium (K^+), calcium (Ca^{2+}), magnesium (Mg^{2+}), aluminum (Al^{3+}), potential acidity (H + Al), soil organic matter (SOM) and pH analyses were determined according to Raij et al. (2001) and cation exchange capacity (CEC) calculated based on the contents of these elements.

Particle size analysis (sand, silt and clay), macroporosity (macro), microporosity (micro) and total pore volume (TPV)

were determined according to Embrapa (1997) and soil bulk density (Ds) using underformed core samples. The total pore volume (TPV) was calculated from Ds, with pore size distribution determined by a porous plate funnel under a 60-cm water column tension in previously saturated samples. Air-filled pore space (AFPS) was calculated as the difference between the porosity fraction filled with water (Ms) and TPV. Organic carbon and nitrogen content were determined using the methods described by Tedesco et al. (1995) and Bataglia et al. (1983), respectively. Carbon stock (Cstock) and nitrogen stock (Nstock) were calculated based on the equivalent soil mass (Carvalho et al. 2009), while the humification index of soil organic matter (H_{LIFS}) was determined using laser-induced fluorescence spectroscopy (LIFS; Milori et al. 2006). In this study, we used the interaction of humification index organic matter and soil bulk density ($H_{LIFSxDs}$) since it better expresses the amount of humic acid at each sampling point. The preparation procedures and soil sample analysis were conducted according to the methods described by Santos et al. (2015).

Machine learning (ML) models

Machine learning (ML) is an effective empirical approach for both regression and/or classification usually applied to complex nonlinear problems with high variability in space and time (Zendehboudi et al. 2018; Hamrani et al. 2020). In this approach, input data is provided for the algorithms to learn from experience and predict a variable of interest (Abbasi et al. 2021). In this study, we have used shallow neural network (SNN), classical regression and hybrid model to predict soil CO_2 emission from reforested areas. The ML models we used are explained below.

Generalized regression neural network (GRNN)

GRNN is based on a probability density function with four layers, one being the input, two hidden and one output (Specht et al. 1991). The main advantages of this typology can be summarized as follows: simplicity of structure; fast approximation procedures and it avoids falling into the local minimum problem because the transfer function is the Gaussian radial basis function (Freitas et al. 2018; Zhang et al. 2022). The number of neurons in the first layer was: 250 (reforested eucalyptus, pine and native species) and 750 (generalized scenario); while the number of neurons in the second layer was 2 for all scenarios.

Radial basis function neural network (RBFNN)

RBFNN is based on supervised learning and is similar to neural network feed-forward neural network (FNN) with radial basis functions as activation functions (Abbasi et al. 2021). This typology presents a simple structure in terms of the information flow direction: an input layer, a hidden layer and an output layer (Han et al. 2011). In this study we used the Gaussian function because it possesses only a few hyperparameters (Hamrani et al. 2020). The number of neurons in the first layer was: 29 (eucalyptus), 65 (pine), 74 (native species) and 100 (generalized scenario) and this topology has only one hidden layer, while the activation function is the radial basis function, defined by unsupervised techniques (Haykin 2001).

Multilayer perceptron neural network (MLPNN)

MLPNN is composed of more than one perceptron and trained by the backpropagation (BP) algorithm, which is responsible for performing weight and bias adjustments in relation to the error, which can be measured in several ways (Haykin 2001). These algorithms are often applied to supervised learning problems: they train on a set of input–output pairs and learn to model the correlation between these inputs and outputs (Freitas et al. 2018). For this topology, the number of neurons in the first layer was: 8 (eucalyptus), 11 (pine), 6 (native species) and 10 (generalized scenario); while the number of neurons in the second layer was 1 for all scenarios. The activation function used was sigmoid (or logistics).

Adaptative neuro-fuzzy inference system (ANFIS)

The ANFIS model is trained by a hybrid learning algorithm consisting of a combination of gradient descent and the least squares method (Jang 1993; Fig. S1). This model combines the learning resources of an artificial neural network with fuzzy logic to provide advanced prediction resources and make the approach more robust than it would normally be with only one of these techniques (Yilmaz and Kaynar 2011). It is important to underscore that in the adaptative neuro-fuzzy inference system training process, the number of input variables may hinder the execution and applicability of the system, since the number of fuzzy rules and computational time increase exponentially if the number of inputs is greater than five (Kaab et al. 2019). In this study, the hyperparameters were: MF type: gaussian function; number labels: 5; max. iterations: 2 (eucalyptus), 6 (pine), 10 (native species) and 5 (generalized scenario); step size: 0.1; total number of parameters: 10 (eucalyptus, pine and native species) e 13 (generalized scenario). An explanation of the mathematical model and the functions of each layer of this model are given in Table S1.

Random forest (RF)

RF is a supervised ML classifier based on decision trees (Breiman 2001). These decision trees use bootstrap aggregating called "bagging" and from the original data they generate a bootstrap sample, and train a model using this bootstrap data (Khaledian and Miller 2020). The random

forest algorithm has been applied for forecasting in several studies, such as the ones that have been done on GHG emissions (Tavares et al. 2018; Hamrani et al. 2020; Abbasi et al. 2021; Kumari and Singh 2022; Singh et al. 2022). The hyperparameters used were: eucalyptus (number estimators: 400; random state: 8; max. depth: 10; min. impurity decrease: 0; min. samples split: 2), pine (number estimators: 200; random state: 10; max. depth: 12; min. impurity decrease: 0; min. samples split: 2), native species (number estimators: 300; random state: 10; max. depth: 15; min. impurity decrease: 0; min. samples split: 2), generalized scenario (number estimators: 200; random state: 10; max. depth: 16; min. impurity decrease: 0; min. samples split: 2). No normalization/standardization procedure was used.

Data pre-processing

Initially, the dataset consisted of 32 variables, including chemical, physical and climatic attributes (Table S2). However, a preliminary analysis was conducted to reduce the number of variables and determine which input parameters have the highest statistical significance on the prediction of soil CO_2 emissions in reforested areas. This analysis involved: Pearson's correlation at a 5% significance level, canonical correlation analysis (CCA) and biophysical justification. After this preliminary analysis, the descriptive statistics (mean, standard error and coefficient of variation) were calculated for each quantitative variable (Table 1). Subsequently, was assessment/removal of multivariate outliers using the Mahalanobis distance at a 1% significance level and tested the hypothesis of multivariate normality.

In addition to decreasing input parameters, the CCA was applied to investigate the associations between chemical and physical properties and climatic attributes (Cruz and Regazzi 1994). These associations are analyzed using canonical variables, which are constructed from the linear combinations of two groups of variables. In this study, the first group was represented for soil carbon dynamics (U), and the second group was formed by soil physicochemical properties and climatic variables represented by (V). In the present study, the soil carbon

Table 1 Descriptive statistics of soil CO_2 emission, soil temperature, soil moisture, soil chemical and physical attributes and climate variables in reforested areas of eucalyptus, pine and native species. Source: Vicentini et al. 2019 (Adapted)

	Eucalyptus			Pine			Native Species		
Attributes	Mean	SE	CV	Mean	SE	CV	Mean	SE	CV
FCO_2	5.61a	0.095	37.81	4.06b	0.069	38.26	5.53a	0.083	33.44
Ts	26.52a	0.046	3.84	25.99a	0.058	4.99	25.77a	0.059	5.13
Ms	10.67b	0.197	42.22	11.57b	0.170	32.95	15.62a	0.234	33.43
pH	4.31a	0.017	8.89	4.00b	0.005	2.96	4.44a	0.013	6.42
Al	5.88b	0.158	60.01	13.92a	0.160	25.72	5.08b	0.194	85.24
H + Al	55.68b	0.573	23.01	74.80a	0.855	25.57	48.80b	0.413	18.94
P	6.52a	0.113	38.87	6.32a	0.112	39.75	6.16a	0.102	36.96
SB	31.32b	0.823	58.76	9.12c	0.160	39.21	34.18a	0.636	41.60
CEC	87.00a	0.680	17.49	83.92a	0.797	21.24	82.98a	0.386	10.41
$H_{LIFSxDS}$	54,713.41b	577.202	23.59	74,913.21a	1041.645	31.09	50,348.11b	329.316	14.63
Cstock	20.67a	0.172	18.60	13.92c	0.086	13.88	17.50b	0.090	11.51
Nstock	1.47a	0.021	32.37	0.90b	0.004	8.98	1.57a	0.007	10.22
Macro	8.17a	0.145	39.71	4.27b	0.087	45.67	5.29b	0.160	67.64
Micro	31.02b	0.150	10.79	36.84b	0.099	5.99	39.37a	0.143	8.12
AFPS	28.52a	0.268	21.04	29.54a	0.207	15.69	29.05b	0.271	20.89
Sand	61.09b	0.217	7.93	64.83a	0.115	3.97	54.24c	0.114	4.70
Silt	5.65a	0.037	14.80	2.58b	0.080	69.79	6.52a	0.061	20.84
Clay	33.26b	0.206	13.84	32.60b	0.143	9.81	39.24a	0.132	7.51
Tair	27.04	0.062	5.16	27.04	0.062	5.16	27.04	0.062	5.16
Humidity	77.84	0.294	8.44	77.84	0.294	8.44	77.84	0.294	8.44
Precipitation	2.76	0.138	111.22	2.76	0.138	111.22	2.76	0.138	111.22

N: 25; means followed by the same letter (lower case) in the columns do not differ (Tukey's test; $p < 0.05$)

FCO_2: soil CO_2 emission (μmol m^{-2} s^{-1}); Ts: soil temperature (ºC); Ms: soil moisture (%); Al: aluminium content ($mmol_c$ dm^{-3}); H + Al: potential acidity ($mmol_c$ dm^{-3}); P: phosphorous (mg dm^{-3}); SB: sum of bases (mmolc dm^{-3}); CEC: cation exchange capacity ($mmol_c$ dm^{-3}); $H_{LIFSxDS}$: interaction between the interaction of humification index and soil bulk density (arbitrary unit); Cstock: soil carbon stock (mg ha^{-1}); Nstock: soil nitrogen stock (mg ha^{-1}); Macro: soil macroporosity (%); Micro: soil microporosity (%); AFPS: air-filled pore space (%); sand, silt and clay contents (%); Tair: air temperature (ºC); Humidity: relative humidity (%); Precipitation: precipitation (mm day^{-1})

dynamics (U) are: FCO_2, carbon stock and $H_{LIFSxDS,}$ while physicochemical and climate variables correspond to second group (V): Ts, Ms, pH, Al, P, CEC, nitrogen stock, macro, micro, sand, silt, clay, Tair, humidity and precipitation.

After the preliminary analysis, the input parameters decreased to 20 independent variables: Ts, Ms, soil chemical (pH, Al, H + Al, P, SB, CEC, $H_{LIFSxDS}$, carbon stock, nitrogen stock), physical properties (macro, micro, AFPS, sand, silt, clay), climate attributes (Tair, humidity and precipitation) and one dependent variable: FCO_2.

Reforested scenarios for modelling

Four reforestation scenarios: reforested eucalyptus (RE); reforested pine (RP); reforested native species (RNS) and a combination with the entire dataset: generalized scenario (GS) data were investigated in this study. Each reforestation scenario has 500 observations measured in the field and the generalized scenario therefore presents 1500 observations to each variable. Of these data, 75% were used for training and 25% to test the models. The criterion to interrupt the calibration algorithm used was a total number of 50 cycles (Haykin 2001). Data pre-processing and ML models were implemented in Python and R environment (R Core Team 2022).

Model performance metrics

The predictive performance of the five ML models (GRNN, RBFNN, MLPNN, ANFIS and RF) applied in this study were analysed using various statistical metrics. The following metrics were calculated: mean absolute error (MAE), root mean squared error (RMSE); mean absolute percentage error (MAPE), Pearson's correlation coefficient (r), adjusted coefficient of determination (R^2 adj), index of agreement (d) (Willmott 1981) and confidence coefficient (c). The equations of the metric used in this study to evaluate the predictive performance of the models are given in Table S3. The analysis of variance of the five ML models was conducted by the F-test (probability of 0.01). Methodology flowchart can be seen in Fig. 2.

Results

Input variable selection

A preliminary analysis was conducted to reduce the initial amount of input parameters and identify the properties that most influence the soil CO_2 emission (FCO_2) in tropical forests. This analysis involved: Pearson's correlation, canonical correlation and biophysical justification. Figure 3 shows Pearson's correlation for the four reforestation scenarios studied: reforested eucalyptus (RE), reforested pine (RP), reforested native species (RNS) and generalized scenario (GS). For the RE, RP and RNS scenarios we used fewer input variables (10) than the generalized scenario (13). GS was composed of the combination of three different ecosystems and is therefore highly heterogeneous in terms of nutrient cycling, soil carbon dynamics, quantity and quality of material deposited on the surface of the soil. Therefore, generalized scenario needed more input variables than the other models for better predictive performance. Overall, the variables that showed the highest correlation with FCO_2 were: P content, $H_{LIFSxDS}$, Cstock, Nstock, microporosity and precipitation (Fig. 3a-d).

Canonical correlation analysis (CCA) was used to investigate the associations between chemical and physical properties and climatic attributes and use as a data input selection technique to measure the effect of predictors on soil CO_2 efflux. CCA and its respective canonical R^2 and the significant test are presented in Table 2. Canonical (0.868, 0.587 and 0.357) and R^2 correlations (0.753, 0.344 and 0.128) were significant ($p < 0.01$). Because of the significance of the correlations presented, it can be concluded that the selected groups are not independent. According to the scatterplot of the standardized values of the variables (Fig. 4), these properties show a high positive linear association ($r = 0.87$). It can be inferred that the high values of the canonical variable that presented the physical–chemical and climate attributes (V1) were associated with the high values of the canonical variable that represents the variables related to soil carbon dynamics (U1) since they present a positive linear association.

Table 3 shows the correlations between the variables of the canonical component that represents them (U_k and V_k), called canonical loads as well as the correlations of these variables with another canonical component, known as transversal canonical loads. These values help interpret the canonical variables, since the higher the absolute value of a canonical load, the greater the association between the original variable and the respective canonical component. Therefore, it was observed that the total proportions, which were explained separately by the canonical variables U1 and V1, were equal to 59.56% and 20.43%, respectively. That is, the variable U1 represented 59.56% of the total variation of the group of variables related to soil carbon dynamics attributes, and V1 represented 20.43% of the total variation of the group of physical–chemical variables of the soil and climatic parameters.

The variable FCO_2 presented the highest canonical loading (in module) for U3 (Table 3), therefore, we can understand this component as the emission of CO_2 from the soil, while V3 corresponds to the relationship between FCO_2 and the chemical and physical variables of the soil and climate (Table 3). In order of relevance, the properties that presented the highest canonical loading in V3 were: precipitation (0.475), Ts (0.459), Ms (0.311) and humidity (0.271) (Table 3). Thus, canonical correlation analysis showed that these are the most expressive variables to

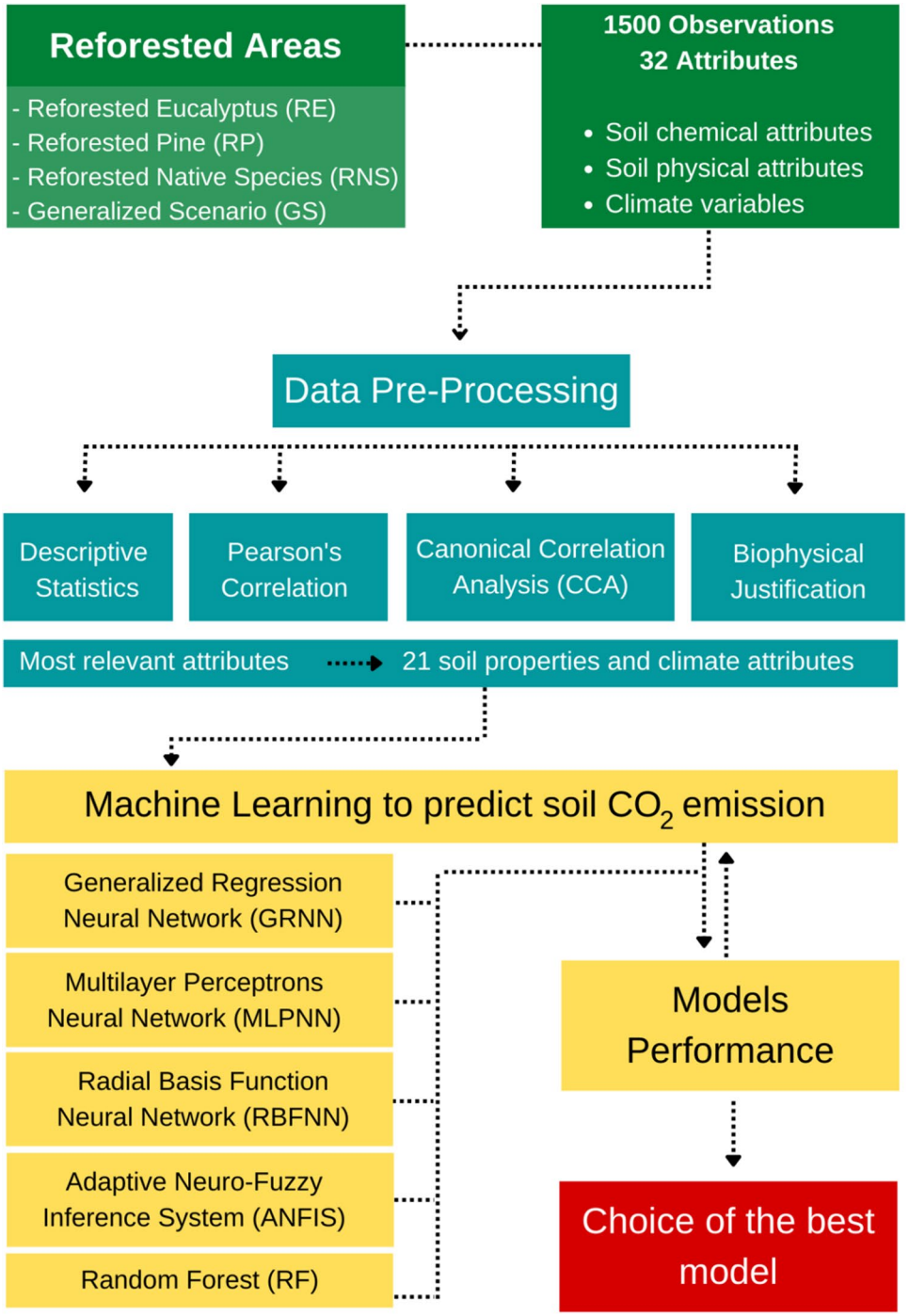

Fig. 2 Methodology flowchart. In green: study areas and input variables; In blue: data pre-processing; In yellow: machine learning models and predictive performance; In red: choice of the best model. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article)

explain the efflux of CO_2 from the soil in tropical forests. We found that the influence of soil and climate attributes were different for each reforestation area, therefore, the input parameters (or predictors) for machine learning models were also different for each ecosystem (Table 4). The attributes Ts, Ms, $H_{LIFSxDS}$ and Cstock were incorporated in the selection of input data for the reforestation scenarios, regardless of the Pearson's or CCA correlation, due to the reported importance of these variables to explain the spatial and temporal variability of soil CO_2 emissions (Silva et al. 2019; Vicentini et al. 2019; Abbasi et al. 2021; Moitinho et al. 2021).

Predictive performance of the ML models

Predictive performance metrics (training and forecasting) of soil CO_2 efflux for reforested eucalyptus, pine, native species and generalized scenario obtained from five ML models are

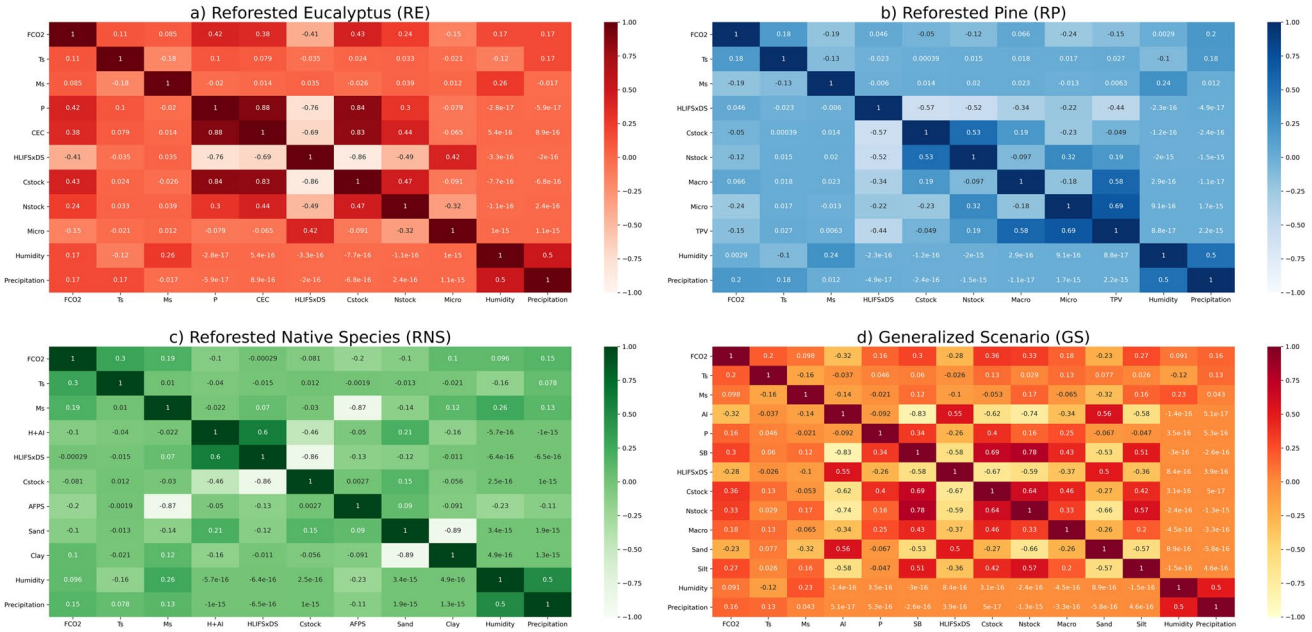

Fig. 3 Correlation matrix between soil physical and chemical attributes and climate: FCO_2: soil CO_2 emission (μmol m^{-2} s^{-1}); Ts: soil temperature (ºC); Ms: soil moisture (%); Al: aluminium content ($mmol_c$ dm^{-3}); H + Al: potential acidity ($mmol_c$ dm^{-3}); P: phosphorous (mg dm^{-3}); SB: sum of bases (mmolc dm^{-3}); CEC: cation exchange capacity ($mmol_c$ dm^{-3}); $H_{LIFSxDS}$: interaction between the interaction of humification index and soil bulk density (arbitrary unit); Cstock: soil carbon stock (mg ha^{-1}); Nstock: soil nitrogen stock (mg ha^{-1}); Macro: soil macroporosity (%); Micro: soil microporosity (%); AFPS: air-filled pore space (%); sand, silt and clay contents (%); Humidity: relative humidity (%); Precipitation: precipitation (mm day^{-1}). Darker color scale indicates positive correlations, while lighter color scale indicates negative correlations

Table 2 Eigenvalues and canonical correlations for soil dynamic, physicochemical and climatic attributes

Pairs of canonical variables[a]	Canonical correlation	Canonical R^2 (eigenvalues)	χ^2	DF[b]	p-value[c]
(U1, V1)	**0.868**	0.753	2915.58	45	< 0.0001**
(U2, V2)	**0.587**	0.344	831.50	28	< 0.0001**
(U3, V3)	**0.357**	0.128	203.57	13	< 0.0001**

[a]U_n = canonical variable for attributes related to soil carbon dynamics, V_n = canonical variable for climate, chemical and physical attributes of the soil, [b]degrees of freedom. [c]**significant at 0.01 probability

Highlighted values represent significant correlations according to the χ^2 test at 5% probability

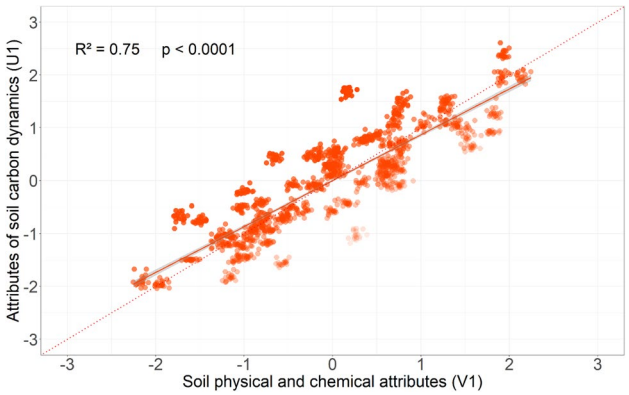

Fig. 4 Scatterplot of the first pair of standardized canonical variables. Attributes of soil carbon dynamics (U1): FCO_2, Cstock, $H_{LIFSxDS}$; physicochemical and climate attributes (V1): Ts, Ms, pH, Al, H + Al, P, CEC, Nstock, macro, micro, AFPS, sand, silt, clay, Tair, humidity and prec

given in Table 5. Figures 5, 6 and 7 show the results of soil CO_2 fluxes for observed and predicted values in the training and prediction stage for RE, RP and RNS, respectively. Heatmaps shows the ranking of machine learning models for the statistical metrics used. It is observed that RF had the best performance in the training phase for all reforestation scenarios: RE (R^2 adj = 0.80, RMSE = 0.93 μmol m^{-2} s^{-1}), RP (R^2 adj = 0.80, RMSE = 0.78 μmol m^{-2} s^{-1}), RNS (R^2 adj = 0.80, RMSE = 0.90 μmol m^{-2} s^{-1}), including the generalized scenario: (R^2 adj = 0.81, RMSE = 0.85 μmol m^{-2} s^{-1}; Fig. 8). Among shallow machine learning models, generalized regression neural network performed better for reforested eucalyptus scenarios (R^2 adj = 0.73, RMSE = 1.02 μmol m^{-2} s^{-1}; Fig. 5), reforested native species (R^2 adj = 0.70, RMSE = 0.94 μmol m^{-2} s^{-1}; Fig. 7), and generalized scenario (R^2 adj = 0.74, RMSE = 0.91 μmol m^{-2} s^{-1}; Fig. 8), while for reforested pine, the radial basis function neural network model pointed out the best metrics (R^2

Table 3 Correlations between soil carbon dynamics and climate, chemical and physical attributes of the soil, based on canonical correlation analysis (CCA)

Dependent Attributes	Canonical loads			Canonical cross-loading		
	U1	U2	U3	V1	V2	V3
FCO_2	0.388	-0.006	0.320	0.447	-0.010	0.895
Cstock	0.851	0.106	-0.030	0.980	0.180	-0.083
$H_{LIFSxDS}$	-0.687	0.354	0.032	-0.792	0.604	0.089
PVE (%)	59.56	13.26	27.18	-	-	-
RI	52.90	-	-	-	-	-
Independent Attributes	Canonical loads			Canonical cross-loading		
	V1	V2	V3	U1	U2	U3
Ts	0.143	0.163	0.459	0.124	0.096	0.164
Ms	-0.009	-0.341	0.311	-0.008	-0.200	0.111
pH	0.641	-0.089	-0.159	0.557	-0.052	-0.057
Al	-0.742	0.122	-0.108	-0.644	0.072	-0.039
P	0.452	0.136	-0.046	0.392	0.080	-0.017
CEC	0.578	-0.226	-0.285	0.502	-0.133	-0.102
Nstock	0.773	-0.183	0.106	0.671	-0.108	0.038
Macro	0.535	0.015	-0.097	0.464	0.009	-0.035
Micro	-0.415	-0.614	0.006	-0.360	-0.360	0.002
Sand	-0.405	0.647	-0.219	-0.351	0.379	-0.078
Silt	0.512	-0.042	0.231	0.444	-0.025	0.083
Clay	0.243	-0.748	0.153	0.211	-0.439	0.055
Tair	-0.007	0.004	-0.197	-0.006	0.002	-0.070
Humidity	0.010	-0.005	0.271	0.009	-0.003	0.097
Precipitation	0.018	-0.009	0.475	0.015	-0.005	0.170
PVE (%)	20.43	10.84	6.08	-	-	-
RI	19.89	-	-	-	-	-

FCO_2: soil CO_2 emission; Cstock: soil carbon stock; $H_{LIFSxDS}$: interaction of humification index of soil organic matter with soil bulk density; Ts: soil temperature; Ms: soil moisture; Al: aluminium content; P: phosphorous content; CEC: cation exchange capacity; Nstock: soil nitrogen stock; Macro: soil macroporosity; Micro: soil microporosity; Tair: air temperature; Humidity: relative humidity; PVE (%): proportion of variation explained; RI: redundancy index

Table 4 Input and number of variables for each reforestation scenario studied

Reforested areas	Input variables	Amount
Reforested Eucalyptus (RE)	Ts, Ms, P, CEC, $H_{LIFSxDS}$, Cstock, Nstock, micro, humidity and precipitation	10
Reforested Pine (RP)	Ts, Ms, $H_{LIFSxDS}$, Cstock, Nstock, micro, macro, TPV, humidity and precipitation	10
Reforested Native Species (RNS)	Ts, Ms, H + Al, $H_{LIFSxDS}$, Cstock, AFPS, sand, clay, humidity and precipitation	10
Generalized Scenario (GS)	Ts, Ms, Al, P, SB, $H_{LIFSxDS}$, Cstock, Nstock, macro, sand, silt, humidity and precipitation	13

adj = 0.53, RMSE = 0.97 μmol m^{-2} s^{-1}; Fig. 6). For all scenarios, adaptative neuro-fuzzy inference system showed the worst predictive performance, with R^2 adj between 0.22 and 0.38 and RMSE ranging from 1.40 to 1.63 μmol m^{-2} s^{-1} (Table 5; Figs. 5, 6, 7 and 8). Apart from ANFIS, most machine learning models in the training phase showed a high correlation ($r \geq 0.64$) and the indices of agreement were above 0.74 (Willmott 1981), demonstrating a very good fit between the observed and estimated values in the respective ML models, according to the classification index (Camargo and Sentelhas 1997; Table 5).

In the predicting stage, random forest indicated the best metrics for reforested eucalyptus, pine and generalized scenario, with error indices and correlations being similar for all the models, exhibiting low RMSE (1.02 to 1.07 μmol m^{-2} s^{-1}) and MAE (0.79 to 0.82 μmol m^{-2} s^{-1}) and correlations varying between 0.70 and 0.84. The reforested native species scenario had the best results for generalized regression neural network (R^2 adj = 0.64, RMSE = 1.04 μmol m^{-2} s^{-1}). Radial basis function neural network ranked 2nd for the three reforested scenarios (eucalyptus, pine and native species). Similar to the

Table 5 Comparative assessment of the predictive performance of soil CO_2 emissions for the training and predicting stages using machine learning models and biophysical models from reforested eucalyptus (RE), reforested pine (RP), reforested native species (RNS) and generalized scenario (GS)

Topology	Phase	MAE	RMSE	MAPE	r	R^2 adj	d	Classification (c)		F	p-value
Reforested Eucalyptus (RE) performance											
GRNN 10–250-2[a]	Training	0.80	1.02	16.14	0.86	0.73	0.92	0.79	Very good	976.20	< 0.0001
	Predicting	0.93	1.14	19.04	0.77	0.59	0.87	0.67	Good	161.90	< 0.0001
RBFNN 10–29-1	Training	0.86	1.09	16.93	0.84	0.71	0.91	0.77	Very good	868.70	< 0.0001
	Predicting	0.87	1.09	17.93	0.83	0.68	0.90	0.75	Good	250.90	< 0.0001
MLPNN 10–8-1	Training	0.88	1.09	18.03	0.82	0.68	0.90	0.74	Good	742.00	< 0.0001
	Predicting	0.94	1.11	19.45	0.78	0.60	0.87	0.67	Good	174.50	< 0.0001
ANFIS	Training	1.15	1.41	25.94	0.62	0.38	0.72	0.44	Poor	185.50	< 0.0001
	Predicting	1.16	1.39	24.98	0.59	0.34	0.72	0.42	Poor	73.89	< 0.0001
RF	Training	0.70	0.93	14.04	0.90	0.80	0.93	0.83	Very good	1408.00	< 0.0001
	Predicting	0.79	1.02	16.28	0.84	0.70	0.91	0.75	Good	320.70	< 0.0001
Reforested Pine (RP) performance											
GRNN 10–250-2	Training	0.77	1.01	22.24	0.74	0.55	0.81	0.60	Passable	447.80	< 0.0001
	Predicting	0.88	1.12	25.34	0.62	0.37	0.72	0.44	Poor	80.28	< 0.0001
RBFNN 10–65-1	Training	0.73	0.97	20.93	0.73	0.53	0.83	0.60	Passable	404.40	< 0.0001
	Predicting	0.83	1.08	24.74	0.64	0.40	0.75	0.48	Poor	87.62	< 0.0001
MLPNN 10–11-1	Training	0.85	1.08	23.98	0.64	0.40	0.74	0.47	Poor	245.70	< 0.0001
	Predicting	0.94	1.14	25.76	0.60	0.36	0.67	0.40	Very poor	66.32	< 0.0001
ANFIS	Training	1.17	1.41	33.50	0.47	0.22	0.70	0.32	Very poor	74.80	< 0.0001
	Predicting	1.25	1.49	36.63	0.40	0.16	0.65	0.26	Very poor	22.38	< 0.0001
RF	Training	0.58	0.78	16.59	0.89	0.80	0.90	0.81	Very good	1400.00	< 0.0001
	Predicting	0.80	1.07	20.91	0.70	0.48	0.81	0.57	Passable	108.40	< 0.0001
Reforested Native Species (RNS) performance											
GRNN 10–250-2	Training	0.71	0.94	14.43	0.84	0.70	0.90	0.75	Good	853.10	< 0.0001
	Predicting	0.85	1.04	17.41	0.80	0.64	0.87	0.70	Good	207.60	< 0.0001
RBFNN 10–74-1	Training	0.87	1.10	17.68	0.78	0.61	0.88	0.68	Good	563.50	< 0.0001
	Predicting	0.96	1.19	19.60	0.77	0.59	0.87	0.67	Good	166.20	< 0.0001
MLPNN 10–6-1	Training	0.88	1.12	18.43	0.70	0.49	0.81	0.57	Passable	345.90	< 0.0001
	Predicting	0.96	1.25	20.77	0.66	0.43	0.77	0.51	Passable	95.12	< 0.0001
ANFIS	Training	1.41	1.63	29.66	0.52	0.27	0.67	0.35	Very poor	106.10	< 0.0001
	Predicting	1.39	1.60	30.41	0.49	0.23	0.60	0.29	Very poor	37.44	< 0.0001
RF	Training	0.67	0.90	13.15	0.90	0.80	0.90	0.81	Very good	1398.00	< 0.0001
	Predicting	0.96	1.18	19.44	0.67	0.45	0.78	0.52	Passable	106.80	< 0.0001
Generalized Scenario (GS) performance											
GRNN 13–750-2	Training	0.70	0.91	15.64	0.86	0.74	0.92	0.71	Good	853.10	< 0.0001
	Predicting	0.82	1.05	20.16	0.81	0.65	0.89	0.71	Good	207.60	< 0.0001
RBFNN 13–100-1	Training	0.81	1.03	18.56	0.82	0.67	0.90	0.74	Good	563.50	< 0.0001
	Predicting	0.90	1.11	20.95	0.77	0.60	0.87	0.68	Good	166.20	< 0.0001
MLPNN 13–10-1	Training	0.93	1.17	21.91	0.75	0.57	0.85	0.64	Average	345.90	< 0.0001
	Predicting	0.96	1.19	23.20	0.75	0.55	0.83	0.62	Average	95.12	< 0.0001
ANFIS	Training	1.16	1.40	28.36	0.55	0.30	0.71	0.39	Very poor	106.10	< 0.0001
	Predicting	1.19	1.44	29.17	0.52	0.27	0.68	0.35	Very poor	37.44	< 0.0001
RF	Training	0.64	0.85	14.81	0.90	0.81	0.93	0.83	Very good	1398.00	< 0.0001
	Predicting	0.80	1.01	18.49	0.84	0.70	0.89	0.75	Good	106.80	< 0.0001

[a]Sequence of numbers indicates the number of input variables, number of neurons in the first layer and number of neurons in the second layer, respectively; MAE: mean absolute error; RMSE: root mean squared error; MAPE: mean absolute percentage error; r = Pearson's correlation; R^2 adj: adjusted coefficient of determination; d = Willmott's index of agreement; c = confidence coefficient; p-value = probability of the occurrence of H_0

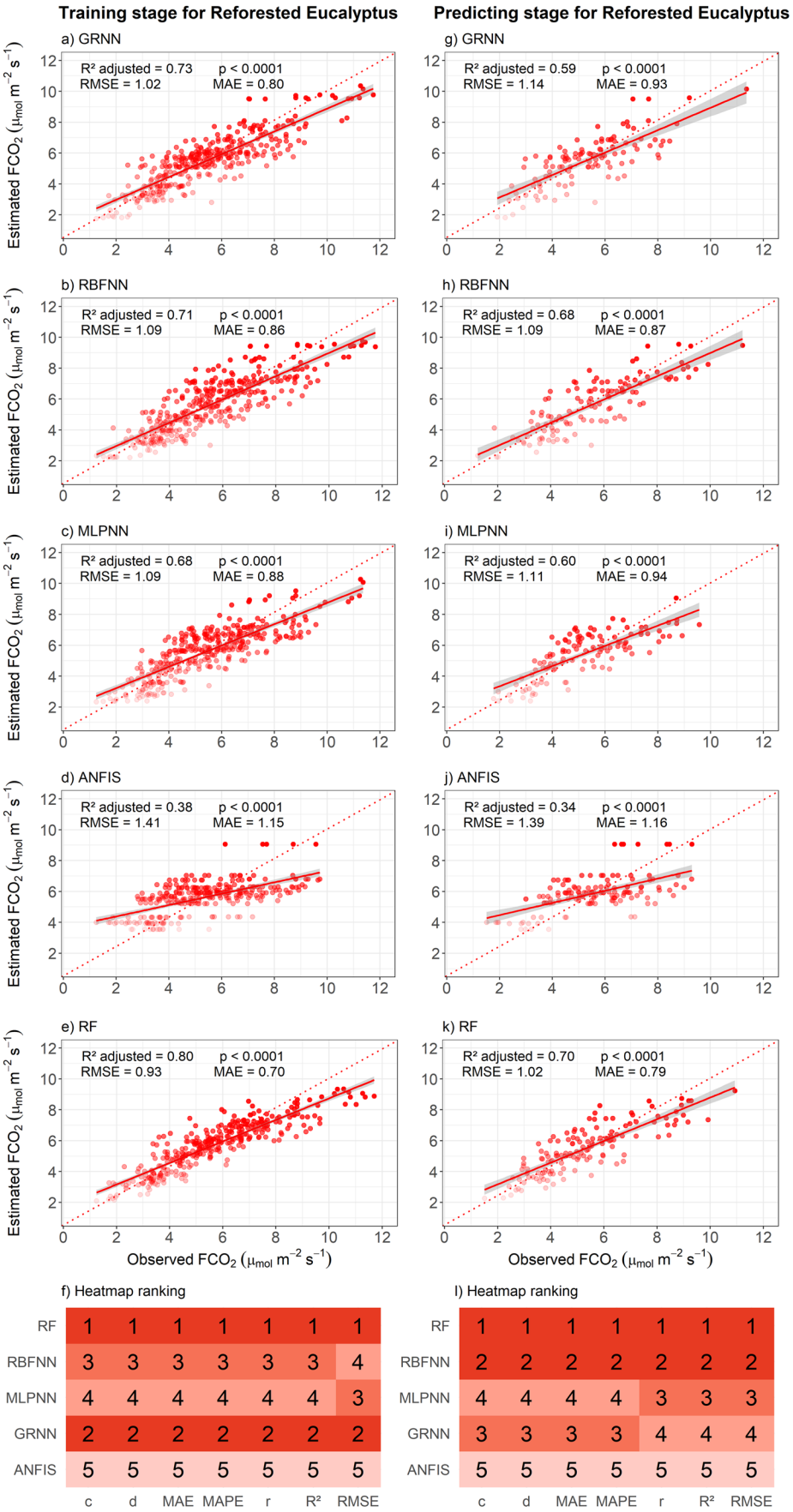

Fig. 5 Soil CO_2 emission observed vs predicted (performance graph) in the training and predicting stage for reforested eucalyptus (RE) for the machine learning models. **a**) GRNN: generalized regression neural network; **b**) RBFNN: radial basis function neural network; **c**) MLPNN: multilayer perceptron neural network; **d**) ANFIS: adaptive neuro-fuzzy inference system; **e**) RF: random forest; **f**) heatmap ranking according to predictive performance

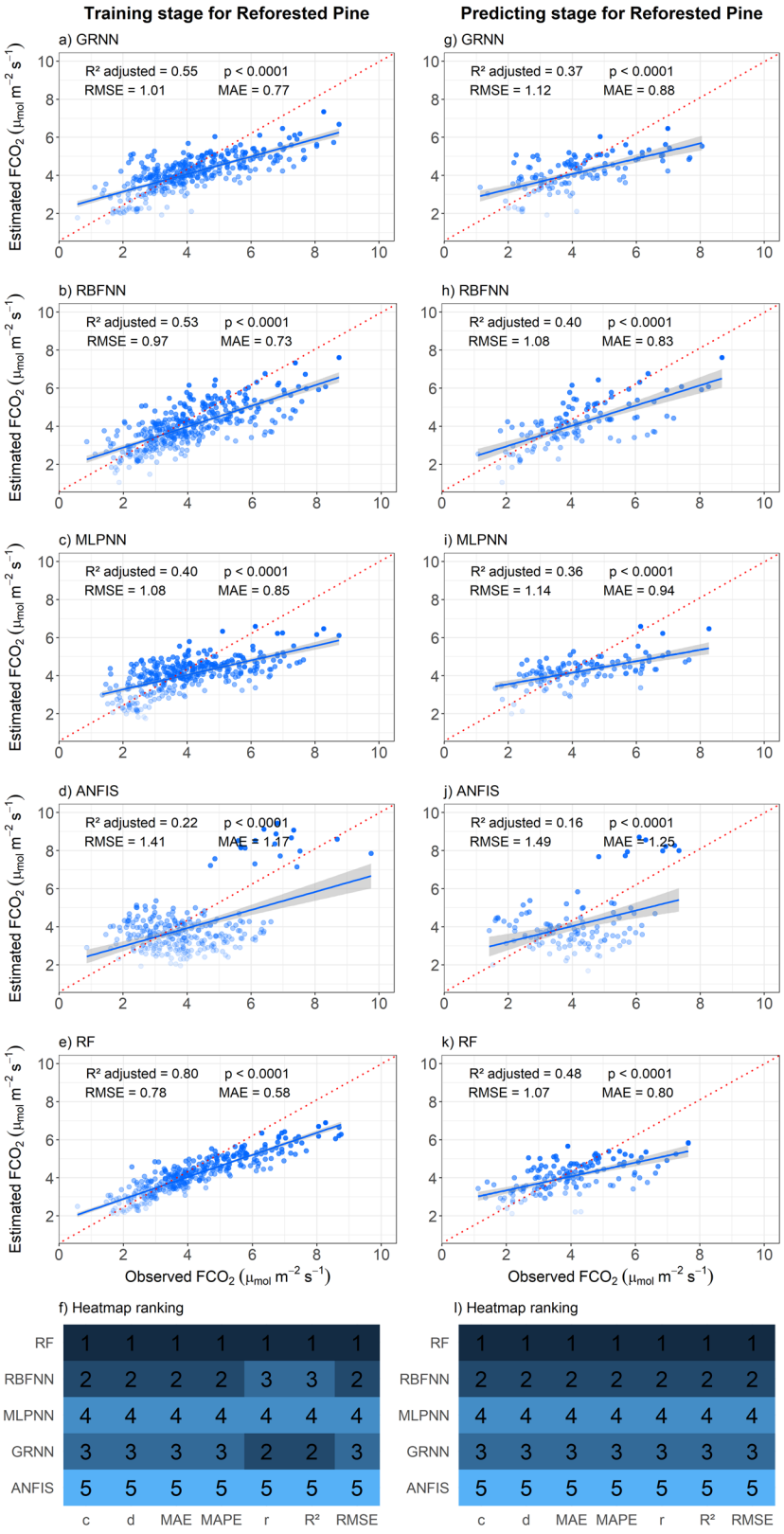

Fig. 6 Soil CO_2 emission observed vs predicted (performance graph) in the training and predicting stage for reforested pine (RP) for the machine learning models. **a**) GRNN: generalized regression neural network; **b**) RBFNN: radial basis function neural network; **c**) MLPNN: multilayer perceptron neural network; **d**) ANFIS: adaptive neuro-fuzzy inference system; **e**) RF: random forest; **f**) heatmap ranking according to predictive performance

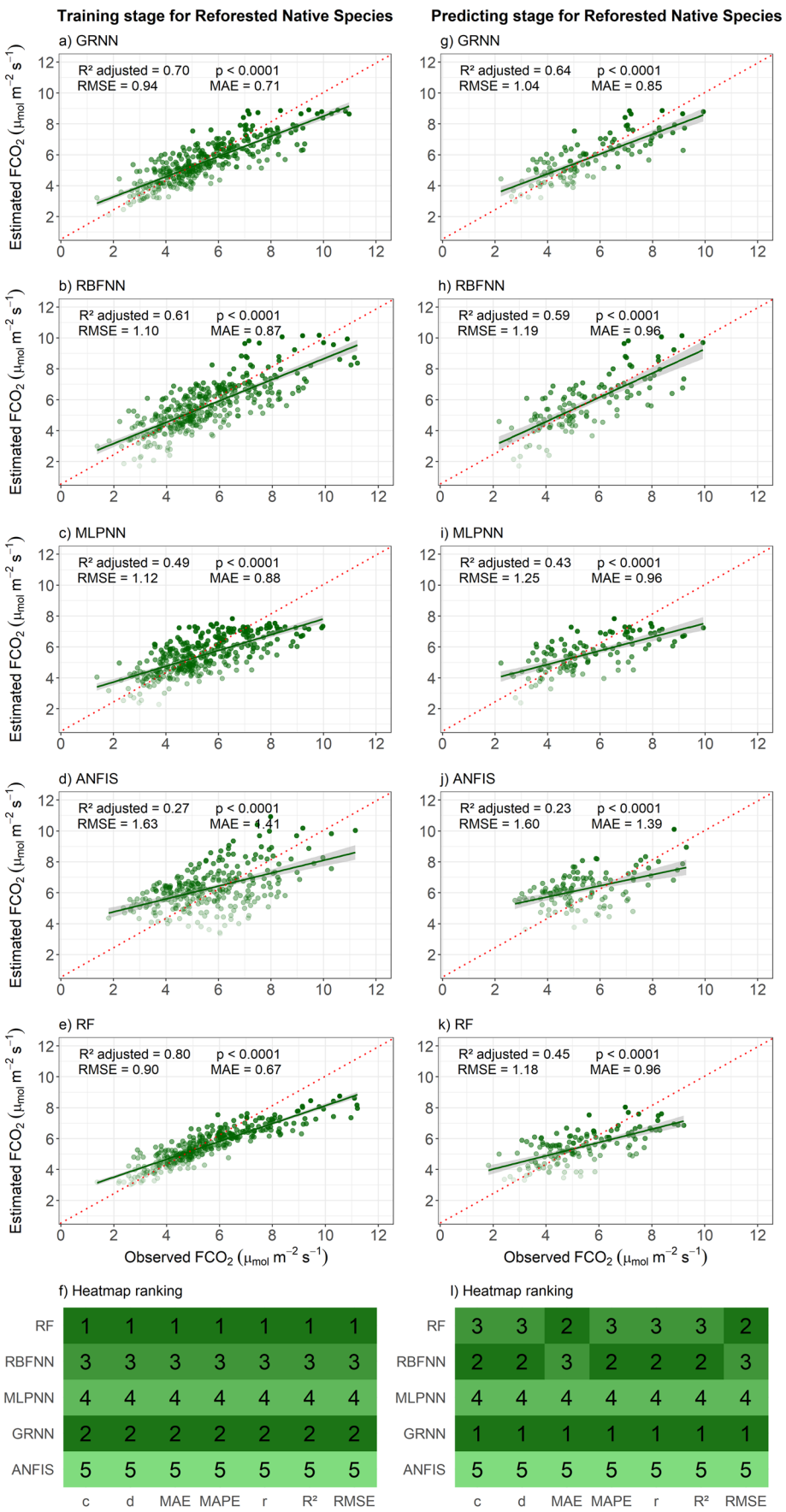

Fig. 7 Soil CO_2 emission observed vs predicted (performance graph) in the training and predicting stage for reforested native species (RNS) for the machine learning models. **a**) GRNN: generalized regression neural network; **b**) RBFNN: radial basis function neural network; **c**) MLPNN: multilayer perceptron neural network; **d**) ANFIS: adaptive neuro-fuzzy inference system; **e**) RF: random forest; **f**) heatmap ranking according to predictive performance

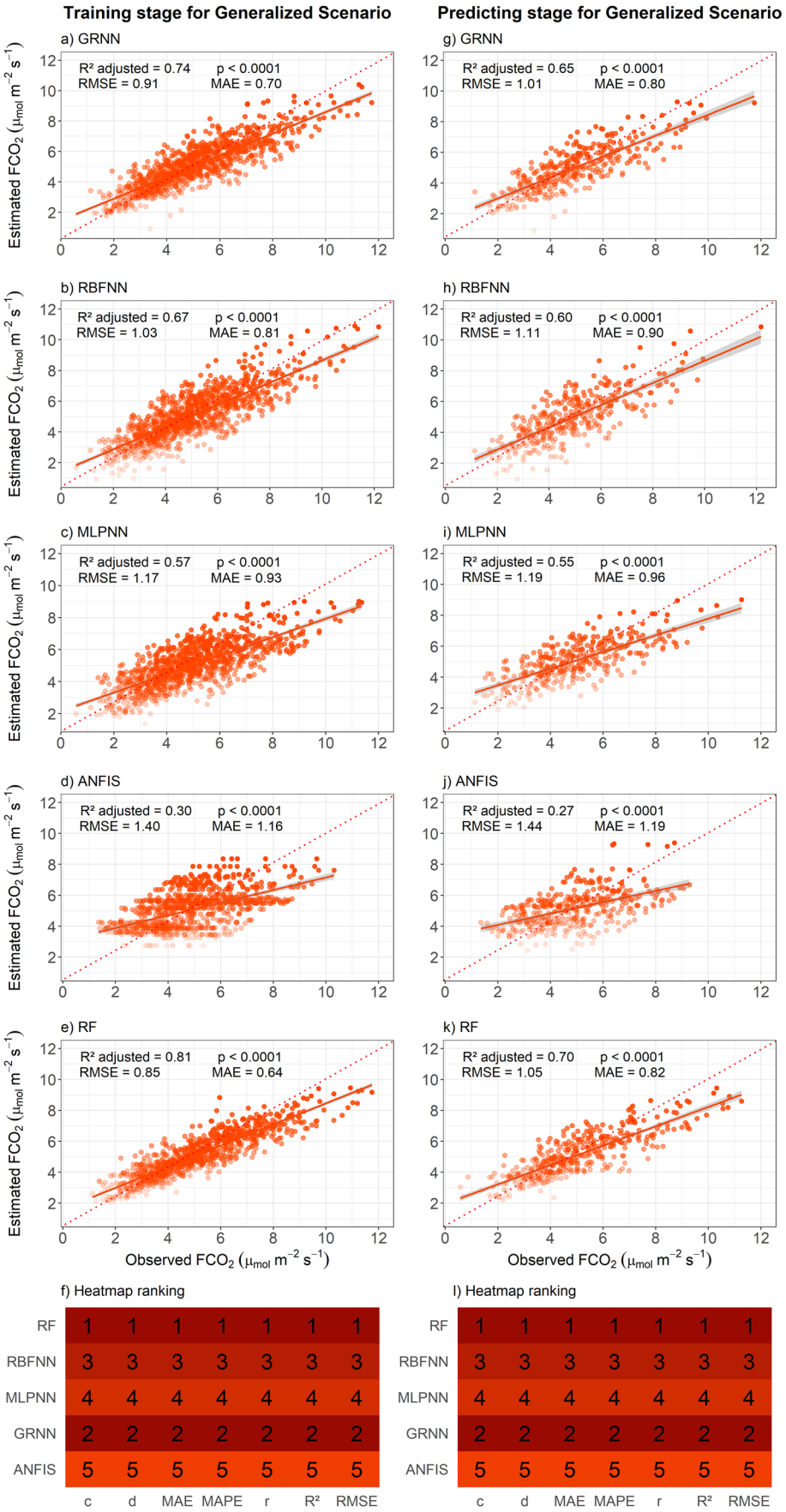

Fig. 8 Soil CO_2 emission observed vs predicted (performance graph) in the training and predicting stage for generalized scenario (GS) for the machine learning models. **a**) GRNN: generalized regression neural network; **b**) RBFNN: radial basis function neural network; **c**) MLPNN: multilayer perceptron neural network; **d**) ANFIS: adaptive neuro-fuzzy inference system; **e**) RF: random forest; **f**) heatmap ranking according to predictive performance

training phase, ANFIS remained the worst machine learning model, showing high error rates (RMSE = 1.39 to 1.60 μmol m^{-2} s^{-1}, MAE = 1.16 to 1.39 μmol m^{-2} s^{-1}) and R^2 adj between 0.16 to 0.34. From the heatmap ranking of the prediction performances with regards to all metrics for the scenarios reforested eucalyptus and pine, the ML models can be classified from best to worst, as follows: (1) RF, (2) RBFNN, (3) GRNN, (4) MLPNN, (5) ANFIS. For reforested native species, the ranking is as follows: (1) GRNN, (2) RBFNN, (3) RF, (4) MLPNN, (5) ANFIS. Finally, for generalized scenario, the ranking is as follows: (1) RF, (2) GRNN, (3) RBFNN, (4) MLPNN, (5) ANFIS.

Discussion

Input variable selection

In this study, chemical and physical parameters of soil and climate were selected based on Pearson's correlation, canonical correlation and biophysical justification between the analyzed variables and the efflux of CO_2 from the soil. The selection of input variables is an indispensable step to optimize the predictive performance of ML models (Noori et al. 2010). The effect and importance of soil and climate properties on soil CO_2 emissions has been well documented (Tavares et al. 2016; Silva et al. 2019; Vicentini et al. 2019; Moitinho et al. 2021). In general, the most expressive variables according to Pearson's correlation were: P content, $H_{LIFSxDS}$, Cstock, Nstock, microporosity and precipitation. The available phosphorus content is associated with the nutritional status and microbial activity of the soil (Moitinho et al. 2021). Therefore, this element can be a limiting factor for the development of microbial communities, since it participates in the dynamics and intensity of the metabolism of microorganisms (Duah-Yentumi et al. 1998). Spohn and Schleuss (2019) reported that the addition of inorganic P to soils in forest systems increased microbial respiration, this is because the added inorganic phosphorus exchanges with organic compounds sorbed in the soil and thus makes these compounds available for microbial decomposition. The authors suggest that such mechanisms are strongly applied to soils rich in Fe and Al oxides and hydroxides, such as the tropical soils in this study.

The degree of humification of organic matter (H_{LIFS}) reflects the quality of the fractions (labile or recalcitrant) of carbon associated with the organic material and, therefore, constitutes an important indicator of the quality of organic matter (Bordonal et al. 2017). Low levels of H_{LIFS} indicate the predominance of labile forms of carbon, which are available and easily assimilated by microbiological communities. On the other hand, more humidified soil carbon fractions are more resistant to microbial activity (Vicentini et al. 2019). This attribute is particularly important, as the ecosystems described in this study accumulate and make available organic compounds in different quantities and qualities and, therefore, the dynamics in the decomposition rates of these materials is influenced by the microenvironment (Xiao et al. 2019). In addition, organic matter is the primary source of energy for microorganisms to use in their metabolic functions and consequent production of CO_2 from the soil (Vicentini et al. 2019). Therefore, the cycling of nutrients and soil carbon dynamics in tropical planted forests (eucalyptus, pine and reforestation with native species) indicate the complexity of the sequences and chemical and physical transformations that occur in these environments.

In general, carbon stock obtained the highest correlations with the soil physicochemical component, given that adequate soil management, by providing a greater amount and cycling of organic matter, may significantly maximize carbon stock over time (Souza et al. 2018). Li et al. (2012) observed that the carbon stock dynamics are closely related to nitrogen stock dynamics, because the increase in nitrogen stock reduces the limitation of nitrogen and contributes to long-term carbon sequestration. Thus, if carbon sequestration is not accompanied by a simultaneous gain in nitrogen, the system will suffer an imbalance, thereby reducing the carbon sequestration rate and intensifying soil CO_2 emission.

Soil structural characteristics represent another important component linked to soil respiration, because greater soil aeration promotes greater diffusion of gases from the soil interior to the atmosphere, boosting soil CO_2 emission (Wick et al. 2012; Tavares et al. 2016). In a study carried out in three contrasting cropping systems, Silva et al. (2019) observed that the distribution of pore sizes and classes influence water storage and drainage, which in turn regulate the movement of gases in the soil and directly affect the soil CO_2 efflux. The authors reported the highest soil CO_2 emissions associated with macropore classes and low FCO_2 for the no-till system with a predominance of micropore classes. According to Fick's gas diffusion law, macro and micropores present antagonistic behavior regarding soil respiration. Soils with higher levels of macropores offer a less tortuous route for CO_2 molecules, facilitating the flow of gases. On the other hand, microporosity provides a more irregular path, which makes gas exchange difficult and reduces CO_2 emission from the soil (Tavares et al. 2015).

The production of CO_2 results from a biochemical process related to the respiration of plant roots and, therefore, is highly influenced by climatic factors, such as precipitation (Ussiri and Lal 2009), which in turn regulates soil moisture (Abbasi et al. 2021). In particular, Ts and Ms are the main drivers of the temporal variability of FCO_2, as they promote greater soil biological activity (Carbonell-Bojollo et al. 2012; Adjuik and Davis 2022). Soil temperature acts in important physiological (evapotranspiration) and

biochemical (increased enzyme activity) processes. As such, it has been reported as one of the main environmental factors governing daytime respiration, as it regulates the speed of chemical reactions and, therefore, conditions the rapid or slow degradation of organic matter (Wallenstein et al. 2010).

Predictive performance of the ML models

Previous research has also reported promising results when using random forest to model environmental phenomena. In a study on the spatial and temporal prediction of global heterotrophic respiration, Tang et al. (2020) observed that RF had R^2 with a value of 0.50 using climate and soil predictors. Abbasi et al. (2021) studied three nitrogen fertilization scenarios for modeling soil CO_2 emission under corn and soybean rotation and among the six evaluated models, random forest showed the best predictive performances with R^2 between 0.86 and 0.94 and MAPE ranging from 21.09 at 34.98 kg ha^{-1} day^{-1}. In an experiment conducted in areas of raw and burned sugarcane in the state of São Paulo, Tavares et al. (2018) used RF for the prediction of FCO_2 in these systems and had a great fit (R^2 0.80). Likewise, Hamrani et al. (2020) tested the predictive performance of nine ML models to predict FCO_2 in agricultural soils and found random forest as the second-best model, outperforming other classical regression algorithms, shallow neural network (SNN) and even deep learning (DL). Overall, the authors observed that classical regression approaches had the best results when compared with SNN. This fact reveals that the complexity of the model does not always capture the most assertive patterns and suggests that the modeling of cyclic phenomena, such as the emission of CO_2 from the soil, eliminates the need for highly complex models. This has been found to be true in other studies involving GHG prediction (Philibert et al. 2013; Saha et al. 2021).

Generalized regression neural network was the most suitable model for predicting FCO_2 in the reforestation scenario with native species (RSN) and showed a good fit for the other reforestation scenarios. The greatest advantage of GRNN topology is related to its ability to learn the problem more rapidly and thus converge to the optical regression surface with the same speed and a large number of datasets (Pandey and Mishra 2017). In addition, generalized regression neural network provided the highest number of hidden layers to the models. The number of neurons linked to these layers results from the greater complexity of the network with a view to recognizing the nonlinear patterns between the datasets whose variables do not express a direct correlation between them or depend on other multiple variables (Maciel et al. 2012). This trend reveals that forest environments with native species present greater complexity of the factors that regulate soil respiration (possibly due to greater microbiological diversity and nutrient cycling) and suggests that the use of other variables that participate in this process could be used as predictors to improve the predictive performance of models.

According to Kashi et al. (2014), multilayer perceptron neural network models generally have high estimation precision, surpassing radial basis function neural network in many cases. This happens because MLPNN has an architecture with more than one hidden layer and this improves the generalization ability and, consequently, increases the recommendation to use these models compared to RBFNN, for example (Haykin 2001). In contrast, the present study showed that radial basis function neural network was generally ranked as the second-best model in the prediction phase, while multilayer perceptron neural network was ranked fourth. Fernandes et al. (2020) found close predictive performances when using the RBFNN and MLPNN architectures to estimate soil penetration resistance with standardized moisture (R^2: 0.64 and 0.68; RMSE: 0.31 and 0.25 MPa, respectively). When comparing the radial basis function and multilayer perceptron neural networks models to model soil carbon stock in forest ecosystems, Cheshmberah et al. (2020) reported the lowest errors and highest determination coefficient associated with RBFNN. Our findings agree with those of Chen et al. (2018), who used linear regression models (RLM and MNLR) and three different artificial neural networks architectures (multilayer perceptron, backpropagation and generalized regression neural networks) to model the CO_2 emissions of reservoirs. According to the authors, artificial neural networks provided a better predictive capacity than the multiple regression models and the generalized regression neural network obtained the best performance. Using artificial neural networks to predict FCO_2 in sugarcane areas in the state of São Paulo, Freitas et al. (2018) tested different network architectures (GRNN and MLPNN). However, unlike the present study, the best fits ($R^2 = 0.92$) and smallest errors were reported for the multilayer perceptron neural network topology (MAPE = 18.29%), since generalized regression neural network exhibited the highest MAPE (43.02%).

Studies involving the application of artificial neural networks and adaptative neuro-fuzzy inference system to model gases (NO_x, HC, CO, CO_2 and PM) (Gopalakrishnan et al. 2011) and factors that affect yield and cost of biodiesel production (Najafi et al. 2018) showed good predictive performance. Therefore, it is important to underscore that different soil attributes and climatic parameters were more adequate than others as predictive variables in the respective models. This highlights the heterogeneity of the environments studied, given that the attributes exhibit different sensitivities to the forest ecosystems this study. However, according to the statistical indices and depending on generalization, some models are more adequate for determining

datasets than others. Thus, based on the statistical performance of the ML models (Table 5), this study demonstrated that the classical regression RF provides greater precision than GRNN, RBFNN, MLPNN and ANFIS for most forest scenarios in this study. In addition, another important factor in selecting the best models is the similar precision between the calibration and validation phases, since the greater the difference between statistical indices during these phases, the less generalist and more prone to error the model tends to be (Sargent 2013).

Soil carbon provides numerous ecosystem services, such as nutrient regulation and greenhouse gas mitigation, which are key to carbon economies and markets (Grunwald 2022). Soil respiration is a critical component of the global carbon cycle, but it is a source of great uncertainty due to the great spatial and temporal variability, which makes it difficult to quantify it precisely in terms of CO_2 source for the atmosphere. Therefore, improving understanding of carbon storage, stock, and fluxes in soils is particularly important to address soil health, food security, regenerative agriculture, and soil conservation management (Xiong et al. 2014). In this sense, the modeling of soil respiration, especially by machine learning algorithms, can strongly contribute to reducing the uncertainties associated with the sources and sinks of carbon in the soil. Such information may scientifically subsidize the formulation of strategic plans that leverage low-carbon agriculture to mitigate and adapt to global climate change.

Main challenges and limitations this study

Soil CO_2 emission represents an essential component of the carbon cycle. Therefore, modeling this phenomenon and understanding its dynamics, especially in tropical environments, is a strategic action to mitigate GHG emissions. However, the factors that affect the efflux of CO_2 from the soil are spatially and temporally dynamic. Thus, obtaining an ideal parameterization capable of extrapolating the models to very heterogeneous conditions in terms of cultivation practices, seasonal variation, climate and soil is extremely challenging. In addition, our study did not consider variables related to the biological component of the soil, which is considered the most sensitive indicator to environmental changes. Therefore, we strongly encourage future studies to incorporate these variables in the input selection of models that aim to estimate soil CO_2 efflux.

Knowledge about the causal relationship between the processes that regulate soil CO_2 emissions is essential to achieve interpretability and extrapolate the findings to other studies, using the machine learning approach. Therefore, even using algorithms that capture patterns through learning, the pre-selection and pre-processing of pedologically relevant environmental variables in terms of soil formation processes is important from an integrated scientific point of view. In this regard, some researchers have warned about results and knowledge extracted purely from the use of pattern recognition, such as machine learning (McBratney et al. 2019; Reichstein et al. 2019; Wadoux et al. 2020). In summary, our findings support that random forest is a promising tool for estimating soil respiration in tropical forests at a regional and global level, provided that a careful selection of datasets similar to that of the present study is made.

Conclusions

The predictive performance of the four reforested scenarios was highly sensitive to the selection of input variables. An extensive comparison of the models from various statistical metrics revealed that random forest had the best performances in the training and prediction phases for most of the reforestation scenarios. However, it was not robust enough in the prediction phase to predict the CO_2 efflux in a scenario of reforestation with native species. The shallow neural networks (generalized regression and radial basis function neural networks) showed intermediate performances in both phases (training and prediction). The adaptative neuro-fuzzy inference system hybrid model pointed out the worst statistical metrics and therefore we do not recommend it to model FCO_2 in tropical forest.

Our results indicate a potential use of the random forest model to predict cyclical and seasonal trends of CO_2 efflux from soil properties and climatic variables. This approach could contribute to the reduction of uncertainties associated with FCO_2 accountings and help to identify and monitor potential sources and sinks of the main additional greenhouse gas over ecosystems. In addition, this method may be an effective low-cost alternative, especially for underdeveloped countries, since, except for $H_{LIFSxDS}$, most of the model attributes are commonly determined in routine soil analyses.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-023-26824-6.

Acknowledgements We are grateful to the FCAV/UNESP Graduate Soil Science Program for awarding Master's scholarships; the Coordination for the Improvement of Higher Education Personnel (CAPES) for the support (Funding Code 001); the Research Support Foundation of São Paulo State—FAPESP (processes no. 2008/58187-0; 2016/03861-5) and Farm for Teaching, Research and Extension (FEPE/FEIS/UNESP).

Authors contributions **Kleve Freddy Ferreira Canteral:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Maria Elisa Vicentini**: Writing – review & editing, Visualization, Methodology, Investigation, Conceptualization. **Wanderson Benerval De Lucena:** Resources, Investigation. **Mário**

Luiz Teixeira De Moraes: Resources, Investigation. **Rafael Montanari:** Resources, Investigation. **Antonio Sergio Ferraudo:** Resources, Investigation. **Nelson José Peruzzi:** Supervision, Conceptualization. **Newton La Scala Jr:** Writing – review & editing, Supervision, Conceptualization. **Alan Rodrigo Panosso:** Writing – review & editing, Supervision, Conceptualization.

Funding We are grateful to the FCAV/UNESP Graduate Soil Science Program for awarding Master's scholarships; the Coordination for the Improvement of Higher Education Personnel (CAPES) for the support (Funding Code 001); the Research Support Foundation of São Paulo State—FAPESP (processes no. 2008/58187–0; 2016/03861–5) and Farm for Teaching, Research and Extension (FEPE/FEIS/UNESP).

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical approval Not applicable.

Consent to participate All authors consented to participate and contribute to the study.

Consent to publish All authors read and approved the final manuscript for publication.

Competing interests The authors declare that they have no known competing financial interests or personal.

References

Abbasi NA, Hamrani A, Madramootoo CA et al (2021) Modelling carbon dioxide emissions under a maize-soy rotation using machine learning. Biosyst Eng 212:1–18. https://doi.org/10.1016/j.biosystemseng.2021.09.013

Adjuik TA, Davis SC (2022) Machine Learning Approach to Simulate Soil CO_2 Fluxes under Cropping Systems. Agronomy 12:1–18. https://doi.org/10.3390/agronomy12010197

Anache JAA, Flanagan DC, Srivastava A, Wendland EC (2018) Land use and climate change impacts on runoff and soil erosion at the hillslope scale in the Brazilian Cerrado. Sci Total Environ 622–623:140–151. https://doi.org/10.1016/j.scitotenv.2017.11.257

Bataglia OC, Furlani AMC, Teixeira JPF, Furlani PR, Gallo JR (1983) Methods of chemical analysis of plants. Campinas, Boletim Tecnico-Instituto Agronomico (Brazil), p 48, no. 78

Breiman L (2001) Random forests. Mach Learn 45(1):5e32. https://doi.org/10.1023/A:1010933404324

Camargo AP, Sentelhas PC (1997) Avaliação do desempenho de diferentes métodos de estimativa da evapotranspiração potencial no estado de São Paulo. Brasil Rev Bras De Agrometeorol 5(1):89–97

Carbonell-Bojollo RM, Repullo-Ruibérriz De Torres MA, Rodríguez-Lizana A, Ordóñez-Fernández R (2012) Influence of soil and climate conditions on CO_2 emissions from agricultural soils. Water Air Soil Pollut 223:3425–3435. https://doi.org/10.1007/s11270-012-1121-9

Carvalho JLN, Cerri CEP, Feigl BJ et al (2009) Carbon sequestration in agricultural soils in the Cerrado region of the Brazilian Amazon. Soil Tillage Res 103:342–349. https://doi.org/10.1016/j.still.2008.10.022

Cerri CEP, Cerri CC, Maia SMF et al (2018) Reducing Amazon deforestation through agricultural intensification in the Cerrado for advancing food security and mitigating climate change. Sustain 10:1–18. https://doi.org/10.3390/su10040989

Chen Z, Ye X, Huang P (2018) Estimating carbon dioxide (CO_2) emissions from reservoirs using Artificial Neural Networks. Water (Switzerland) 10. https://doi.org/10.3390/w10010026

Cheshmberah F, Fathizad H, Parad GA, Shojaeifar S (2020) Comparison of RBF and MLP neural network performance and regression analysis to estimate carbon sequestration. Int J Environ Sci Technol 17:3891–3900. https://doi.org/10.1007/s13762-020-02696-y

Cruz CD, Regazzi AJ (1994) Modelos Biométricos Aplicados Ao Melhoramento Genético. Universidade Federal de Viçosa, Viçosa

de Bordonal RO, Lal R, Ronquim CC et al (2017) Changes in quantity and quality of soil carbon due to the land-use conversion to sugarcane (Saccharum officinarum) plantation in southern Brazil. Agric Ecosyst Environ 240:54–65. https://doi.org/10.1016/j.agee.2017.02.016

de Silva BO, Moitinho MR, de Santos GAA et al (2019) Soil CO_2 emission and short-term soil pore class distribution after tillage operations. Soil Tillage Res 186:224–232. https://doi.org/10.1016/j.still.2018.10.019

dos Maciel LS, Ballini R, da Silveira RLF (2012) Apreçamento de opções sobre taxa de câmbio R$/US$ negociadas no Brasil: uma comparação entre os modelos Black e redes neurais artificiais. Rev Adm 47:96–111. https://doi.org/10.5700/rausp1028

Dos Santos CH, Romano RA, Nicolodelli G et al (2015) Performance evaluation of a portable laser-induced fluorescence spectroscopy system for the assessment of the humification degree of the soil organic matter. J Braz Chem Soc 26:775–783. https://doi.org/10.5935/0103-5053.20150039

Duah-Yentumi S, Rønn R, Christensen S (1998) Nutrients limiting microbial growth in a tropical forest soil of Ghana under different management. Appl Soil Ecol 8:19–24. https://doi.org/10.1016/S0929-1393(97)00070-X

Embrapa – Empresa Brasileira de Pesquisa Agropecuária (1997) Embrapa – Empresa Brasileira de Pesquisa Agropecuária, 1997. Centro Nacional de Pesquisa de Solos. Manual de métodos de análise de solo. 2nd ed. Ministério da Agricultura e do Abastecimento, Brasília, p 212. (In Portuguese)

Farhate CVV, De Souza ZM, De Medeiros Oliveira SR et al (2018) Use of data mining techniques to classify soil CO_2 emission induced by crop management in sugarcane field. PLoS ONE 13:1–18. https://doi.org/10.1371/journal.pone.0193537

Fernandes MMH, Coelho AP, da Silva MF et al (2020) Estimation of soil penetration resistance with standardized moisture using modeling by artificial neural networks. Catena 189:104505. https://doi.org/10.1016/j.catena.2020.104505

Freitas LPS, Lopes MLM, Carvalho LB et al (2018) Forecasting the spatiotemporal variability of soil CO_2 emissions in sugarcane areas in southeastern Brazil using artificial neural networks. Environ Monit Assess 190. https://doi.org/10.1007/s10661-018-7118-0

Gopalakrishnan K, Mudgal A, Hallmark S (2011) Neuro-fuzzy approach to predictive modeling of emissions from biodiesel powered transit buses. Transport 26:344–352. https://doi.org/10.3846/16484142.2011.634080

Grunwald S (2022) Artificial intelligence and soil carbon modeling demystified: power, potentials, and perils. Carbon Footprints 1:5. https://doi.org/10.20517/cf.2022.03

Hamrani A, Akbarzadeh A, Madramootoo CA (2020) Machine learning for predicting greenhouse gas emissions from agricultural soils. Sci Total Environ 741:140338. https://doi.org/10.1016/j.scitotenv.2020.140338

Han HG, Chen QL, Qiao JF (2011) An efficient self-organizing RBF neural network for water quality prediction. Neural Netw 24(7):717–725. https://doi.org/10.1016/j.neunet.2011.04.006

Haykin S (2001) Redes neurais: princípios e prática. Bookman Editora. https://doi.org/10.1002/0471221546

IPCC – Intergovernmental Panel on Climate Change (2021) Climate Change: Mitigation. Contribution of Working Group III. Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press: Cambridge, United Kingdom and New York, 2021. Available online: https://www.ipcc.ch/working-group/wg3/. Accessed 7 Jun 2022

Jang JR (1993) ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans Syst Man Cybern 23(3):665–685. https://doi.org/10.1109/21.256541

Kaab A, Sharifi M, Mobli H et al (2019) Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production. Sci Total Environ 664:1005–1019. https://doi.org/10.1016/j.scitotenv.2019.02.004

Kashi H, Emamgholizadeh S, Ghorbani H (2014) Estimation of Soil Infiltration and Cation Exchange Capacity Based on Multiple Regression, ANN (RBF, MLP), and ANFIS Models. Commun Soil Sci Plant Anal 45:1195–1213. https://doi.org/10.1080/00103624.2013.874029

Khaledian Y, Miller BA (2020) Selecting appropriate machine learning methods for digital soil mapping. Appl Math Model 81:401–418. https://doi.org/10.1016/j.apm.2019.12.016

Khan MZ, Khan MF (2019) Application of ANFIS, ANN and fuzzy time series models to CO_2 emission from the energy sector and global temperature increase. Int J Clim Chang Strateg Manag 11:622–642. https://doi.org/10.1108/IJCCSM-01-2019-0001

Kumari S, Singh SK (2022) Machine learning-based time series models for effective CO_2 emission prediction in India. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-022-21723-8

Li D, Niu S, Luo Y (2012) Global patterns of the dynamics of soil carbon and nitrogen stocks following afforestation: A meta-analysis. New Phytol 195:172–181. https://doi.org/10.1111/j.1469-8137.2012.04150.x

McBratney A, de Gruijter J, Bryce A (2019) Pedometrics timeline. Geoderma 338:568–575. https://doi.org/10.1016/j.geoderma.2018.11.048

Milori DMBP, Galeti HVA, Martin-Neto L et al (2006) Organic Matter Study of Whole Soil Samples Using Laser-Induced Fluorescence Spectroscopy. Soil Sci Soc Am J 70:57–63. https://doi.org/10.2136/sssaj2004.0270

Moitinho MR, Ferraudo AS, Panosso AR et al (2021) Effects of burned and unburned sugarcane harvesting systems on soil CO_2 emission and soil physical, chemical, and microbiological attributes. Catena 196:104903. https://doi.org/10.1016/j.catena.2020.104903

Najafi B, Faizollahzadeh Ardabili S, Shamshirband S et al (2018) Application of anns, anfis and rsm to estimating and optimizing the parameters that affect the yield and cost of biodiesel production. Eng Appl Comput Fluid Mech 12:611–624. https://doi.org/10.1080/19942060.2018.1502688

Noori R, Hoshyaripour G, Ashrafi K, Araabi BN (2010) Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration. Atmos Environ 44:476–482. https://doi.org/10.1016/j.atmosenv.2009.11.005

Pandey A, Mishra A (2017) Application of artificial neural networks in yield prediction of potato crop. Russ Agric Sci 43:266–272. https://doi.org/10.3103/s1068367417030028

Philibert A, Loyce C, Makowski D (2013) Prediction of N_2O emission from local information with Random Forest. Environ Pollut 177:156–163. https://doi.org/10.1016/j.envpol.2013.02.019

R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/. Accessed 13 Aug 2022

Raij BV, Andrade JC, Cantarella H, Quaggio JA (2001) Análise química para avaliação da fertilidade de solos tropicais. Instituto Agronômico, Campinas, p 285. (In Portuguese)

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

Saha D, Basso B, Robertson GP (2021) Machine learning improves predictions of agricultural nitrous oxide (N_2O) emissions from intensively managed cropping systems. Environ Res Lett 16. https://doi.org/10.1088/1748-9326/abd2f3

Sargent RG (2013) Verification and validation of simulation models. J Simul 7:12–24. https://doi.org/10.1057/jos.2012.20

SEEG. Sistema de Estimativas de Emissões e Remoções de Gases de Efeito Estufa (2022). Available In: https://plataforma.seeg.eco.br/total_emission. Accessed on February 16, 2023. (In Portuguese)

Singh PK, Pandey AK, Ahuja S, Kiran R (2022) Multiple forecasting approach: a prediction of CO_2 emission from the paddy crop in India. Environ Sci Pollut Res 29:25461–25472. https://doi.org/10.1007/s11356-021-17487-2

Smith P, Bustamante M, Ahammad H, Clark H, Dong H, Elsiddig EA, Haberl H, Harper R, House J, Jafari M (2014) Agriculture, forestry and other land use (AFOLU), Climate change 2014: mitigation of climate change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press

Soil Survey Staff. USDA NRCS (2014) Keys to soil taxonomy (12th ed.). United States Department of Agriculture, Natural Resources Conservation Service, Washington, DC

Souza LHC, da Matos ES, de Souza Magalhães CA et al (2018) Soil carbon and nitrogen stocks and physical properties under no-till and conventional tillage cotton-based systems in the Brazilian Cerrado. L Degrad Dev 29:3405–3412. https://doi.org/10.1002/ldr.3105

Specht DF et al (1991) A general regression neural network. IEEE Trans Neural Netw 2(6):568–576

Spohn M, Schleuss PM (2019) Addition of inorganic phosphorus to soil leads to desorption of organic compounds and thus to increased soil respiration. Soil Biol Biochem 130:220–226. https://doi.org/10.1016/j.soilbio.2018.12.018

Tang X, Fan S, Du M et al (2020) Spatial and temporal patterns of global soil heterotrophic respiration in terrestrial ecosystems. Earth Syst Sci Data 12:1037–1051. https://doi.org/10.5194/essd-12-1037-2020

Tavares RLM, de Souza ZM, Siqueira DS et al (2015) Soil CO_2 emission in sugarcane management systems. Acta Agric Scand Sect B Soil Plant Sci 65:755–762. https://doi.org/10.1080/09064710.2015.1061048

Tavares RLM, de Souza ZM, La Scala N et al (2016) Spatial and temporal variability of soil CO_2 flux in sugarcane green harvest systems. Rev Bras Cienc Do Solo 40:1–14. https://doi.org/10.1590/18069657rbcs20150252

Tavares RLM, de Oliveira SR, M, De Barros FMM, et al (2018) Prediction of soil CO_2 flux in sugarcane management systems using the random forest approach. Sci Agric 75:281–287. https://doi.org/10.1590/1678-992x-2017-0095

Tedesco MJ, Gianello C, Bissani CA, Bohnen H, Wolkweiss SJ (1995) Análises de solo, plantas e outros materiais, 2nd edn. Universidade Federal do Rio Grande do Sul, Porto Alegre

Tubiello FN, Salvatore M, Cóndor Golec RD, Ferrara A, Rossi S, Biancalani R, Flammin A et al (2014) Agriculture, forestry and other land use emissions by sources and removals by sinks1990–2011 analysis. FAO Statistics Division. Working Paper Series ESS/14-02

UNFCCC (2013) Views on Land Use, Land-use Change and Forestry Issues Referred toin Decision 2/CMP.7, Paragraphs 5e7. Submissions from Parties and Admitted Observer Organizations 12e18 (SBSTA, UNFCCC, 2013). Disponível em:http://go.nature.com/hLAtTN. Acessado em 15.04.21

Ussiri DAN, Lal R (2009) Long-term tillage effects on soil carbon storage and carbon dioxide emissions in continuous corn cropping

system from an alfisol in Ohio. Soil Tillage Res 104:39–47. https://doi.org/10.1016/j.still.2008.11.008

Vicentini ME, Pinotti CR, Hirai WY et al (2019) CO_2 emission and its relation to soil temperature, moisture, and O_2 absorption in the reforested areas of Cerrado biome, Central Brazil. Plant Soil 444:193–211. https://doi.org/10.1007/s11104-019-04262-z

Wadoux AMC, Samuel-Rosa A, Poggio L, Mulder VL (2020) A note on knowledge discovery and machine learning in digital soil mapping. Eur J Soil Sci 71:133–136. https://doi.org/10.1111/ejss.12909

Wallenstein M, Allison SD, Ernakovich J, Steinweg JM, Sinsabaugh R (2010) Controls on the temperature sensitivity of soil enzymes: a key driver of in situ enzyme activity rates. In: Soil enzymology. Springer, Berlin, Heidelberg, pp 245–258. https://doi.org/10.1007/978-3-642-14225-3_13

Wick AF, Phillips RL, Liebig MA et al (2012) Linkages between soil micro-site properties and CO_2 and N_2O emissions during a simulated thaw for a northern prairie Mollisol. Soil Biol Biochem 50:118–125. https://doi.org/10.1016/j.soilbio.2012.03.010

Willmott CJ (1981) On the validation of models. Phys Geogr 2:184–194. https://doi.org/10.1080/02723646.1981.10642213

Xiao W, Chen HYH, Kumar P et al (2019) Multiple interactions between tree composition and diversity and microbial diversity underly litter decomposition. Geoderma 341:161–171. https://doi.org/10.1016/j.geoderma.2019.01.045

Xiong X, Grunwald S, Myers DB, Kim J, Harris WG (2014) Comerford NB. Holistic environmental soil-landscape modeling of soil organic carbon. Environ Model Softw 57:202–215. https://doi.org/10.1016/j.envsoft.2014.03.004

Yilmaz I, Kaynar O (2011) Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. Expert Syst Appl 38:5958–5966. https://doi.org/10.1016/j.eswa.2010.11.027

Zendehboudi S, Rezaei N, Lohi A (2018) Applications of hybrid models in chemical, petroleum, and energy systems: A systematic review. Appl Energy 228:2539–2566. https://doi.org/10.1016/j.apenergy.2018.06.051

Zhang L, Yan W, Liu Y et al (2022) Simulation of soil CO_2 efflux under different hydrothermal conditions based on general regression neural network. Agric For Meteorol 316:108847. https://doi.org/10.1016/j.agrformet.2022.108847