FASE FRONTIERS OF AGRICULTURAL SCIENCE AND ENGINEERING

Engineering Transactions of CAE

RESEARCH ARTICLE

2026 / Volume 13 / Issue 2 / 25663

Predicting nitrous oxide emissions from soil planted to sugarcane under various irrigation regimes using machine learning models

Rafael T. BONATO[1], Lurdineide de A. B. BORGES[2], Arminda M. CARVALHO[2], Alexsandra D. OLIVEIRA[2], Thaís R. SOUSA[3], Maria L. G. RAMOS[3], Walter Q. RIBEIRO JUNIOR[2], Robélio L. MARCHÃO[2], Fernando A. M. SILVA[2], Díbio L. BORGES (✉)[1,4]

1 University of Brasilia, Department of Mechanical Engineering, Brasilia, DF, 70910-900, Brazil.
2 EMBRAPA Cerrados, BR-020, km 18, Planaltina, DF, 73310-970, Brazil.
3 University of Brasilia, Faculty of Agronomy and Veterinary, Brasilia, DF, 70910-900, Brazil.
4 University of Brasilia, Department of Computer Science, Brasilia, DF, 70910-900, Brazil.

KEYWORDS

Brazilian Cerrado, irrigation regimes, machine learning, multilayer perceptron, random forest, sugarcane nitrous oxide emissions

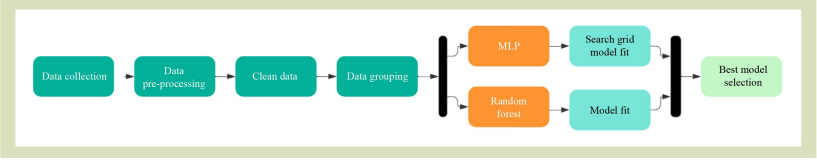
HIGHLIGHTS

- A scalable, data-driven machine learning model for N_2O emissions was refined with data from sugarcane crops.
- A simulation model for forecasting nitrous oxide emissions under various environmental conditions was developed.
- The efficacy of the random forest model for predicting N_2O emissions from irrigated sugarcane fields was determined.
- The relative impacts of watering regimes, soil types and environmental factors on N_2O emissions from sugarcane crops were presented.

Received February 2, 2025;
Accepted July 24, 2025.

Correspondence: dibio@unb.br

GRAPHICAL ABSTRACT



ABSTRACT

Nitrous oxide (N_2O) is a potent greenhouse gas with about 60% of its emissions are attributed to agricultural activities. Its fluxes are influenced by a range of crop-specific factors, such as nitrogenous fertilizer inputs, soil N availability, tillage practices, temperature, pH and soil moisture. These factors interact in complex, nonlinear ways, creating the need for predictive modeling of N_2O emissions to both improve understanding and estimation and identify mitigating strategies. This proposes proposes data-driven machine learning techniques, particularly multilayer perceptron and random forest (RF) algorithms, for estimating soil N_2O fluxes in a sugarcane plantation under different irrigation regimes and to contrast machine learning results with conventional analytical methods. The findings indicate that RF modeling achieved a coefficient of determination of 87.4% for N_2O emission prediction, and identified ammonium, nitrogen nitrate, soil temperature, and water-filled pore space as the most influential predictors, in that order. The results open new possibilities for integrating machine learning to study N_2O fluxes in sugarcane and other major crops. All data and code used in this study are provided openly to support further research.

1 Introduction

Nitrous oxide (N_2O) has a global warming potential about 300 times greater than that of carbon dioxide. It is the third-most significant greenhouse gas in terms radiative forcing and is the leading ozone-depleting substance since its emissions make a direct contribution to stratospheric ozone destruction[1]. About 60% of N_2O emissions originate from human activities, primarily from industrial processes, fossil fuel combustion in large urban centers and agricultural systems[2].

In agriculture, N_2O is mainly produced in the soil by microbial processes of nitrification and denitrification, after nitrogenous fertilizer application[3]. The magnitude of N_2O emissions is strongly influenced by N input levels and soil-climatic factors such as temperature, moisture content, soil type and the availability of key chemical elements. Irrigation is also critical for both in modulating soil N_2O emissions and increasing crop yields[4]. In view of the intensive resource demands of high-yield crops, mainly in terms of land, water and fertilizers, there is an urgent need for analytical tools that will improve understanding of N_2O emission dynamics in these systems[5].

Sugarcane (*Saccharum officinarum*) is a perennial grass native to tropical New Guinea and is grown primarily for sugar and ethanol production. Worldwide, sugarcane plantations cover over 26 Mha. Brazil and India are the largest producers, and together account for more than 50% of global production[6]. Sugarcane production is highly mechanized and demands intensive inputs, particularly regarding water and N fertilizers[7]. A deeper understanding of N_2O emissions from sugarcane fields would be an essential resource to help design effective mitigation strategies to control N_2O emissions from this strategic crop. According to the Intergovernmental Panel on Climate Change[8], the emission factor for croplands, including sugarcane, is estimated at 1%. In other words, for every kilogram of nitrogen fertilizer applied, around 10.5 kg of CO_2 equivalent emissions are released into the atmosphere. Of this total, fertilizer synthesis contributes with about 4.5 kg CO_2 eqv., while transportation, application, and N_2O emissions account for the remaining 6 kg.

N_2O emissions from agricultural soils are influenced by three main categories of factors[9]: environmental, management, and measurement-related variables. All of these components are highly interdependent[10], and substantial efforts have been made to develop process-based modeling tools, for example, DAYCENT[11], DNDC[12] and SWAT[13], to simulate N_2O emissions from agricultural systems. Also, the intricate interactions among these factors, combined with the high spatial-temporal variability of N_2O measurements, pose significant challenges to aggregate and scale emission data[14].

In the Central-West Region of Brazil, within the Cerrado biome, sugarcane production is expanding. Given the prolonged dry periods (usually 160–180 days) in this region, irrigation is a common practice in most cropping areas[15]. Irrigation and fertigation have become increasingly prevalent in sugarcane production in Brazil, particularly as strategies to maintain or increase sugarcane yields under water stress conditions, such as in the Cerrado region. Therefore, irrigation can be applied using different approaches, for example, during crop establishment or in combination with rainfall. Vinasse, a nutrient-rich byproduct of bioethanol distillation, especially high in nitrogen, is frequently used as fertilizer. However, there are growing concerns that its additional application in already fertilized and irrigated areas could exacerbate N_2O emissions. Soil N_2O fluxes from sugarcane areas with N fertilizer and vinasse application are, on average, at least three times greater than when each is applied separately[16]. Recent studies in the Cerrado region confirmed that N fertilizer application in sugarcane fields significantly increases N_2O emissions[16]. Intriguingly, other studies have suggested that under certain irrigation regimes, higher yields can be achieved without proportional increases in N_2O emissions[4]. In light of the increasing availability of experimental data, we hypothesized that data-driven machine learning approaches would provide a new promising framework to predict and test functional and variable relationships, identify critical thresholds and key predictors, and improve overall knowledge within a new paradigm of analysis.

Using data from maize production, Saha et al.[17] demonstrated that machine learning models, particularly random forest (RF), could explain 51% of the variability in daily N_2O fluxes. Their study highlighted the ability of RF models to capture nonlinear relationships among the predictor variables related to N_2O emissions. In a similar study[18], RF modeling was successfully used to examine interactions between predictor variables influencing N_2O emissions from maize and soybean systems. In another study[19], metadata analysis using an RF approach effectively identified significant soil and environmental factors associated with crop residues and N_2O emissions. Also, an optimized irrigation management is a crucial factor that can

enhance crop yields but also potentially mitigate greenhouse gas emissions[4].

In this study, we investigated the main predictor variables of N_2O emissions from sugarcane crops, by integrating different irrigation regimes, soil properties and climatic variables within a nonlinear machine learning framework, specifically multilayer perceptron (MLP) and RF. Although machine learning techniques have been successfully applied previously[17], there is a knowledge gap regarding sugarcane grown in the Cerrado biome or similar conditions, which was addressed by our study. For the analysis, we used experimental data previously examined by standard statistical and principal component analyses[4]. By examining the interactions of these data sets from sugarcane production by applying MLP and RF, we were able to:

(1) assess the relative influence of irrigation regimes, soil types, and environmental variables on N_2O emissions from sugarcane production;

(2) demonstrate the effectiveness of the RF model as the most accurate nonlinear machine learning approach for predicting N_2O emissions from irrigated sugarcane fields;

(3) develop a scalable, data-driven modeling framework for N_2O emissions that can be continuously refined with additional data sets of sugarcane or other crops, to broaden the application spectrum for agricultural emission prediction;

(4) construct a simulation model capable of forecasting N_2O emissions under varying environmental conditions, with a view to providing actionable insights for improved agricultural practices.

2 Materials and methods

2.1 Experimental location

Data were obtained from a study conducted at the experimental station of the Brazilian Agricultural Research Corporation (Embrapa, Cerrados) in Planaltina, DF, Brazil (15°36′18′′S 47°42′36′′W)[4]. The study was installed in the Cerrado biome, where the climate is tropical wet (Aw under Köppen climate classification). The average annual temperature of the microclimate at the study location is between 22 and 23 °C. Most of the 1383 mm of mean annual rainfall occurs between October and March. The soil was categorized as Oxisol with a clay texture. The 0–20 cm soil layer used in this study contained the following chemical properties: pH(H_2O), 5.08; Al^{3+}, 0.39 $cmol_c \cdot dm^{-3}$; P, 0.22 $mg \cdot dm^{-3}$; K^{+}, 8.0 $mg \cdot dm^{-3}$; Ca^{2+}, 0.56 $cmol_c \cdot dm^{-3}$; Mg^{2+}, 0.26 $cmol_c \cdot dm^{-3}$; potential acidity (H^+, Al^{3+}), 3.7 $cmol_c \cdot dm^{-3}$; and organic matter, 8.7 $g \cdot kg^{-1}$.

2.2 Experimental design and measurements

N_2O fluxes were sampled in an area planted to sugarcane in Planaltina, Federal District, Brazil. Immediately after harvest, a uniform initial irrigation of 40 mm was applied across the entire area. The area was then subdivided into 12 plots to implement four distinct irrigation treatments. Each of the three irrigation treatments was applied to three plots, corresponding to 17%, 46% and 75% of the crop evapotranspiration rate (T17%, T46% and T75%, respectively). Three plots received no further irrigation and served as rainfed control (R). The plots were irrigated every 15 days from 23 June to 11 November 2015, totaling 13 irrigations. Table 1 details the irrigation events and the water collected in each treatment (R, T17%, T46% and T75%). During the N_2O sampling period, two N fertilizer applications of 120 kg of ammonium sulfate were applied (22 June and 27 October 2015). As recommended by the Intergovernmental Panel on Climate Change[8], N_2O fluxes were also collected in a native (Cerrado) area adjacent to the experimental field, to establish a baseline reference.

During the monitoring period, 30 N_2O flux collections were made, between 08:30 and 11:30. For sampling, 24 static chambers were installed in the area, two per plot. In each chamber, gas samples were collected at 15 and 30 min after chamber closure at 11:30 to estimate N_2O fluxes. Simultaneously with each gas sampling event, soil temperature (ST) was recorded and soil samples were collected (0–10 cm deep) to determine water-filled pore space (WFPS), ammonium and nitrate concentrations. Table 2 summarizes the measured variables, their acronyms and corresponding units. The complete data set comprises measurements of 13 variables on 30 sampling days, totaling 390 data points to be used in the machine learning experiments for predictive modeling.

Table 1 Amount of water (mm) applied during irrigation events in the experimental area of Embrapa Cerrados[4]

Date of irrigation in 2015	Water collected (mm) in each treatment			
	R	T17%	T46%	T75%
13 June	40.0	40.0	40.0	40.0
23 June	0	6.1	24.6	32.9
2 July	0	3.1	11.4	22.3
13 July	0	4.8	15.5	27.9
23 July	0	6.3	25.7	35.5
3 August	0	7.3	26.1	32.2
12 August	0	4.5	15.1	19.8
24 August	0	7.5	20.9	21.8
3 September	0	6.3	16.2	35.5
14 September	0	8.0	28.6	33.3
24 September	0	3.8	11.9	20.1
5 October	0	6.8	18.7	34.9
15 October	0	2.4	9.2	17.5
11 November	0	2.5	6.8	18.3
Total	40.0	109.0	270.7	392.0

Note: R is the rainfed control with an initial irrigation of 40 mm applied immediately after the previous sugarcane harvest, and T17%, T46% and T75% are irrigation regimes of 17%, 46% and 75% of the crop evapotranspiration rate, respectively.

Table 2 List of experimental variables

Variable	Description
NO_3^-	Nitrate concentration ($mg \cdot kg^{-1}$)
NH_4^+	Ammonium concentration ($mg \cdot kg^{-1}$)
WFPS	Water-filled pore space (%)
ST	Soil temperature (°C)
Tirriga	Irrigation treatment level (% of evapotranspiration)
Precip	Precipitation (mm)
Tmed	Average temperature (°C)
Avg.RH	Average relative humidity (%)
Wind	Wind speed ($m \cdot s^{-1}$)
Insol	Insolation (h)
Rad	Radiation ($MJ \cdot m^{-2} \cdot d^{-1}$)
Evap	Potential evapotranspiration (mm)
N_2O	Nitrous oxide concentration ($\mu g \cdot mg^{-2} \cdot h^{-1}$)

2.3 Machine learning methods

Machine learning is a data-driven computational approach that allows accurate and reliable predictions. It offers powerful tools for exploring and analyzing enormous volumes of complex data sets. These techniques can be classified based on their learning paradigms. In unsupervised learning, algorithms work with unlabeled data to identify patterns or groupings within the data set, whereas in supervised learning, input data are labeled to map the features of an already known target variable. Among the supervised learning approaches used to evaluate performance, RF[20] and the MLP models[21] are widely recognized for their high effectiveness as nonlinear prediction algorithms, which were selected for this study. Figure 1 shows a flowchart of the approach.

2.3.1 Random forest

The RF algorithm, comprising defined number of decision trees (DT), is a highly versatile tool capable of generating multiple solutions to a given problem. Each decision tree is constructed from a root node and expands into sub-trees composed of decision and leaf nodes[20]. This adaptability is evident in the recursive data partitioning process, which forms a hierarchical structure of subsets extending from the root to the leaf nodes. At each node, a splitting function, mostly based

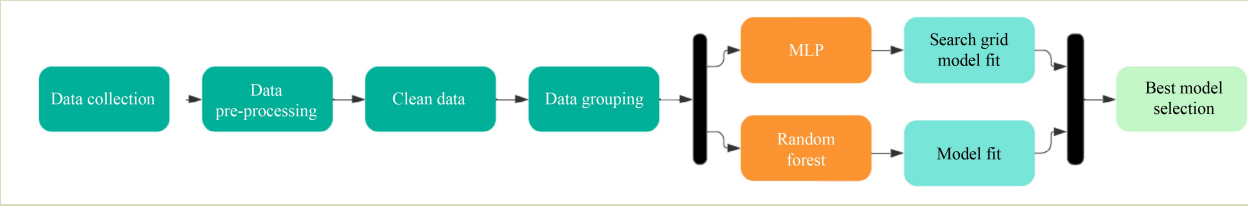


Fig. 1 Flowchart of the proposed approach.

on entropy or Gini impurity, determines the data distribution into more homogeneous subsets. This process continues throughout the data set, progressively defining segmentation intervals at each node until reaching the leaf nodes, where the data subsets are maximally uniform. The subset size represented by each leaf can be a single data point or multiple samples, depending on the depth and complexity of the tree. Tree growth is limited by predefined stopping criteria that govern its structural development.

In this way, the model generates an importance metric for each variable, enabling the assessment of the influence of each input value on the forecast and allowing for the identification of nonlinear interactions among variables. Data are fed into the model in subsets, with repetition in parallel training runs of DT enumerations. Each DT predicts values that are combined in a process called majority voting, where the most commonly predicted outcome is selected as the final prediction. This procedure helps reduce any noise in the data and enhances the robustness of the model. Therefore, Random Forest is widely used in classification and regression problems[22–24].

2.3.2 Multilayer perceptron

The MLP algorithm is an artificial neural network composed of multiple layers of interconnected nodes, or neurons, capable of learning complex data representations[21]. At least three layers are required: an input layer containing the values of the input variables, one or more intermediate hidden layers, and an output layer that provides the predicted classes assigned to the target variable. A series of input values determines the functioning of a neuron; the inputs are used for calculations called activation functions and the result is submitted to an output. These activation functions are nonlinear, enabling the model to represent complex, nonlinear relationships in the data. Neurons in each layer are connected to those in the subsequent layer, with weights assigned to each connection for system adaptation. This neural network is trained with a data set applied to its input layer, and the resulting error values are used by an optimization algorithm in conjunction with the Backpropagation technique which propagates the error backward from the output layer through the network, adjusting the connection weights accordingly[25].

A GridSearch exhaustive search algorithm was applied to optimize the architecture of the perceptron network. This method systematically tests all possible combinations of predefined hyperparameters, and the best parameter configuration is identified according to a specified scoring metric. The parameters established for this simulation included: hidden layer architectures [(50, 50, 5), (100, 100, 5)], activation functions (*relu* and *tanh*); solvers (*adam* and *lbfgs*); learning rate strategies (constant or adaptive), with learning rates from 0.1 to 0.001 of multiples of 10; and the regularization parameter alpha also varied from 0.1 to 0.001 in multiples of 10. The maximum number of training iterations was set to 2000 to ensure efficient evaluation.

2.4 Data preprocessing

The collected data were preprocessed prior to being input into the machine learning algorithms. This process involved filtering to remove missing values and excluding outliers. Due to the substantial differences in variable magnitudes, normalization was performed using min-max scaling. In addition, categorical transformation was applied, separating each treatment and block into independent variables. The analysis was conducted separately for each block, and the average across blocks was used for comparative purposes.

2.5 Model performance metrics

The primary metric used to validate the algorithms was the

mean absolute error (MAE), calculated by averaging the absolute difference between forecast and actual values. The second metric, the mean squared error (MSE), uses the squared value of the MAE metric, which is more sensitive to inaccurate forecasts[25]. The measure of the root mean squared error (RMSE) is derived as the square root of the MSE. Another measurement approach uses the R^2 factor, which indicates the proportion of variance in the observed data that is explained by the model.

$$\mathrm{MAE} = \frac{1}{n} \cdot \sum_{i=0}^{n} |y_i - \hat{y}_i| \quad (1)$$

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 \quad (2)$$

$$\mathrm{RMSE} = c = \sqrt{\frac{1}{n} \cdot \sum_{i=0}^{n} (y_i - \hat{y}_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y_i})^2} \quad (4)$$

where, n is the number of data points, y_i is the actual value for the i-th data point, $\hat{y}_i$ is the predicted value for the i-th data point, and $\overline{y}$ is the mean of the actual values.

2.6 Code and data availability

Code used in this study is available at github[26] which is an open source community. We used the name of github account with the suitable file for code and data

This open access facilitates reproducibility and encourages other researchers to explore new dimensions of the topic in future research.

3 Results and discussion

3.1 Analysis of data variables

The measured N_2O emissions for treatments T75%, T46%, T17%, R and native Cerrado vegetation ranged from −35.7 to 117, −84.6 to 47.8, −26.7 to 92.7, −13.2 to 35.1 and −406 to 61.9 $\mu g \cdot m^{-2} \cdot h^{-1}$ N_2O, respectively. The highest emissions were observed in the T75% treatment and the lowest in the treatments T46% and R. The relationships between irrigation regimes, reference conditions and the variables N_2O, NO_3^-, NH_4^+, and ST are illustrated in Fig. 2, along with their average values. The interactions between the variables were preliminarily assessed through the construction of a correlation matrix (Fig. 3). Significant correlations were primarily associated with climatic parameters. However, interactions involving N_2O emissions showed weak correlations, indicating the absence of a single dominant predictor among the measured variables.

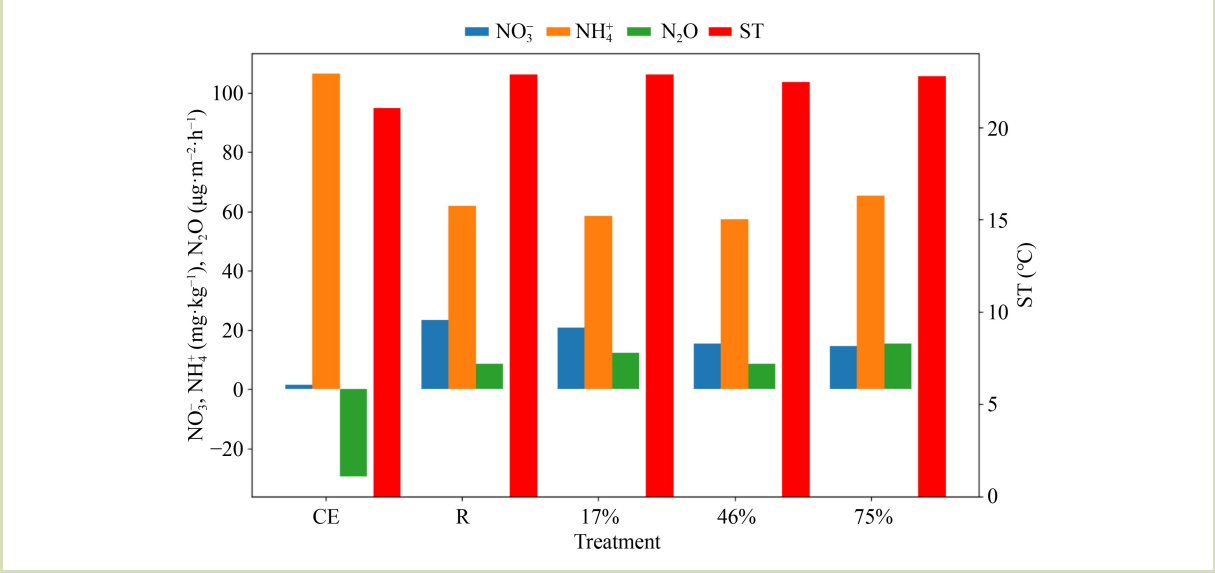


Fig. 2 Plot of average values of N_2O, NO_3^-, NH_4^+ and soil temperature (ST) measured in the treatments: Cerrado (CE), rainfed (R), and irrigation at T17%, T46%, and T75%.

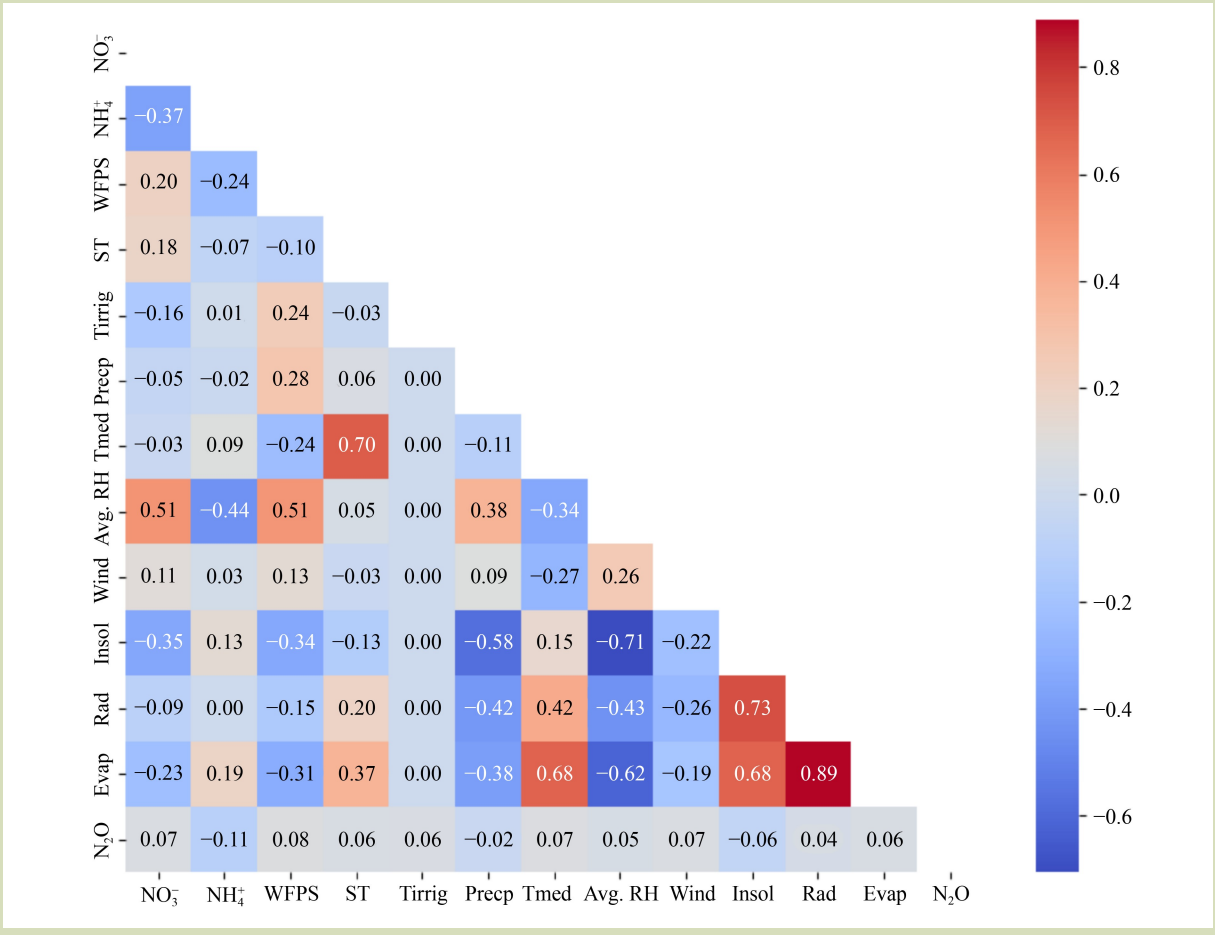


Fig. 3 Correlation matrix for all variables: NO_3^-, NH_4^+, WFPS.

3.2 Machine learning model performance

The N_2O emissions were predicted with two different machine learning models. A RF model was developed based on the averaged values from the three study blocks of soil-, climate-, and gas-related variables. In parallel, an MLP model was trained using the same data with hyperparameter optimization achieved through a systematic search among predefined combinations of parameters to identify the most effective neural network architecture.

Of the 390 measured N_2O emission values, data clustering was observed, that is, a higher concentration of points within certain value intervals. To address this imbalance, two data sets were created for the experiments: a balanced set of data points with 10 subsets of 36 measurements each, evenly distributed across standardized value intervals; and an unbalanced data set containing all original measurements. For both data sets, each machine learning model randomly selected 80% of the points for training and used the remaining unseen 20% for testing. The performance results for RF algorithms are shown in Table 3. The RF model trained on the imbalanced test set obtained the lowest R^2 value (Fig. 4). Conversely, the performance of the balanced data set was superior and able to explain data variation ($R^2 = 0.874$). The MLP algorithm did not perform as well as the RF model (Table 4). Figure 5 illustrates the comparison between observed and predicted N_2O emission values.

Through hyperparameter optimization using the GridSearch algorithm in conjunction with MLP neural networks for predictive modeling, we identified the most influential parameter settings for each data set. For the balanced data set,

Table 3 Metric results of the RF model

Metric	Ungrouped	Unbalanced	Balanced
MAE	4.7914	0.1093	0.3894
MSE	73.0465	0.0462	0.2546
RMSE	8.5467	0.2150	0.5045
R^2	0.8216	0.8094	0.8736

Note: Ungrouped, no partition; imbalanced, class partitions with unequal sizes; balanced, class partitions with equal sizes.

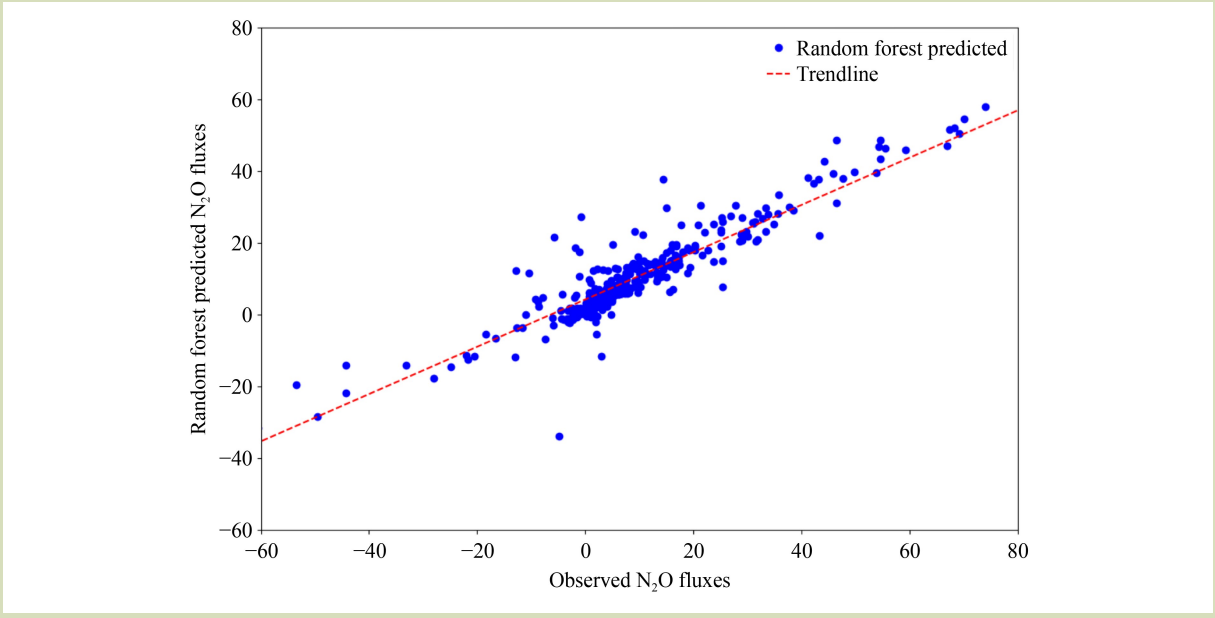


Fig. 4 Observed versus N_2O fluxes predicted by the random forest model.

Table 4 Metric results of the MLP model

Metric	Ungrouped	Unbalanced	Balanced
MAE	1.2574	1.2765	0.5682
MSE	2.2065	2.2320	0.8999
RMSE	1.4854	1.4940	0.9486
R^2	0.3155	0.4055	0.5360

Note: Ungrouped, no partition; imbalanced, class partitions with unequal sizes; balanced, class partitions with equal sizes.

the optimal setup included the *tanh* activation function, hidden layer sizes of 100, 100 and 5 hidden units respectively, a constant learning rate with an initial value of 0.001, a maximum of 2000 iterations, and the *adam* solver. In contrast, the unbalanced data set performed best with the same *tanh* activation function, but required different hyperparameters, namely: hidden layers of 50, 50 and 5 units, respectively, a constant learning rate, an initial learning rate of 0.1, a

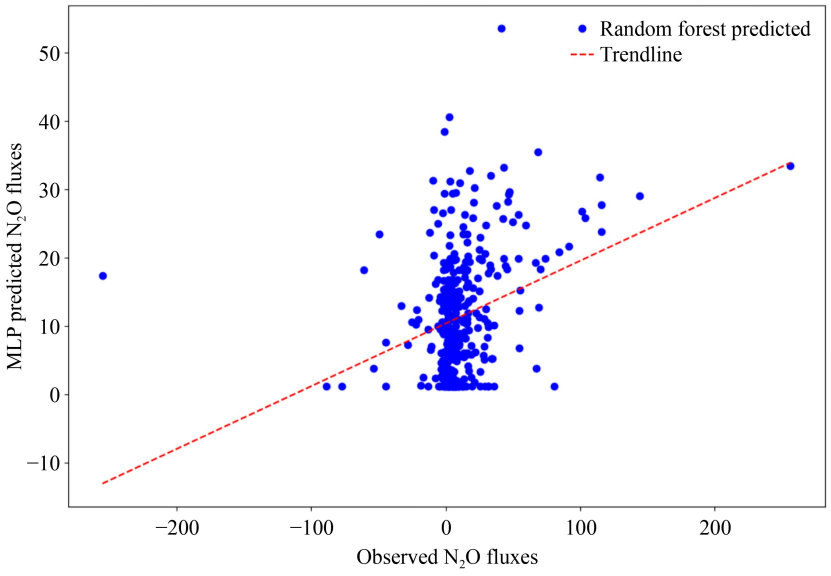


Fig. 5 Observed versus N_2O fluxes predicted by the multilayer perceptron model.

maximum of 2000 iterations, and the *adam* solver. Similarly, the ungrouped data set achieved optimal results with the *tanh* activation function, hidden layers of 50, 50 and 5 neurons, respectively, a constant learning rate, an initial learning rate of 0.1, a maximum of 2000 iterations, and the *adam* solver. This optimized configuration resulted in improved predictive accuracy, highlighting the importance of tailoring hyperparameters to the unique characteristics of each data set.

3.3 Analysis of main variables

Permuted importance measures the percentage increase in model error resulting from the random permutation of a single variable. This approach uses the RMSE to assess how the perturbation in a specific variable affects the average prediction error for N_2O emissions. As shown in Fig. 6, the largest increase in error was associated with two variables related to nitrogen availability in the soil; NH_4^+ as the most important (1.3%) followed by NO_3^- (0.8%). Despite the dominance of these nitrogen-related variables, climate (average RH) and soil characteristics (WFPS and ST) were also among the five most influential of the 12 variables analyzed.

The importance measure permutation importance was used due to its reduced sensitivity to distortions caused by multicollinearity and variable scaling, compared to other importance measures such as Gini importance (Fig. 7). Gini importance, which is based on decision tree splits, tends to overestimate variables with numerous categories. In contrast, permutation importance assesses the effect of each variable directly, independently of variable type or the underlying machine learning algorithm

3.4 Analysis of individual effects of variables

The individual conditional expectation (ICE) is a statistical tool used to evaluate the influence of a specific variable on the output of a predictive model. For each observation in the data set, ICE computes the predicted outcome across a range of values for the variable of interest, while all other variables are held constant, that is, indicating the partial dependence of the response on that variable[27]. The resulting individual prediction curves were plotted in Fig. 8, demonstrating how the model output varies with changes in the selected variable. The permutation importance plot (Fig. 6) identified NH_4^+ as the most important variable. In more detail, the ICE plot for NH_4^+ (Fig. 8) shows a concentration of samples near zero and high emission peaks at 0 and 200, indicating these values as the most significant points for this variable. For NO_3^-, the ICE graph shows considerable variation for values below 13, with

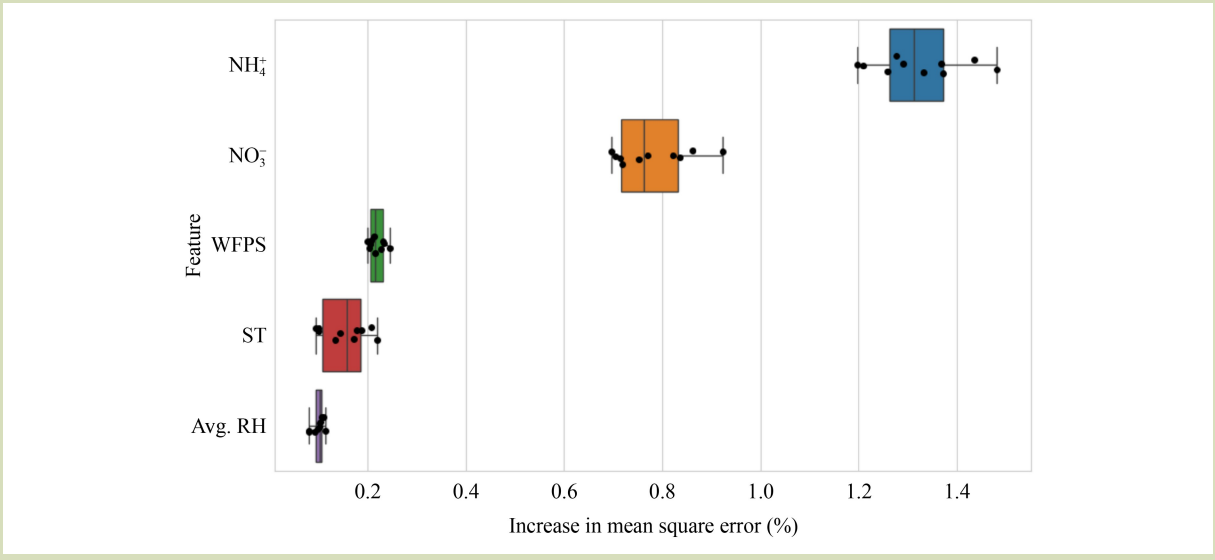


Fig. 6 Boxplots of the distributions of the scaled permutation importance (percentage increase in mean square error) for the top-ranked features: NH_4^+, NO_3^-, water filled pore space (WFPS), soil temperature (ST) and average relative humidity (Avg. RH).

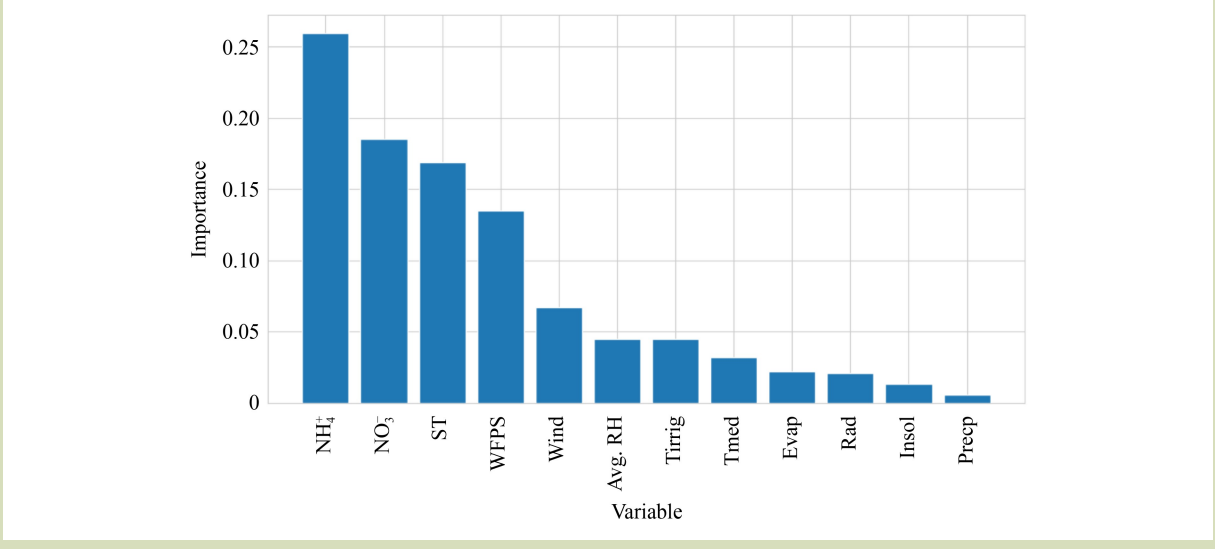


Fig. 7 Scaled variable importance for the measured variables as determined by the random forest model. See Table 2 for variable description.

recurring values ranging from 0 to 80 μg·mg^{-2}·h^{-1} N_2O. Regarding WFPS, the ICE plot shows an increasing trend with a peak at 55%, corresponding to N_2O values between 20 and 50 μg·mg^{-2}·h^{-1}, followed by saturation near 80%. These ICE graphs visually represent the relationships between individual variables and N_2O emissions, highlighting specific patterns of model response to changes in environmental and soil conditions.

The four most significant variables were further analyzed for their contribution to the N_2O emission predictions (Fig. 9). An absence of NH_4^+ accumulation in the soil indicates that N

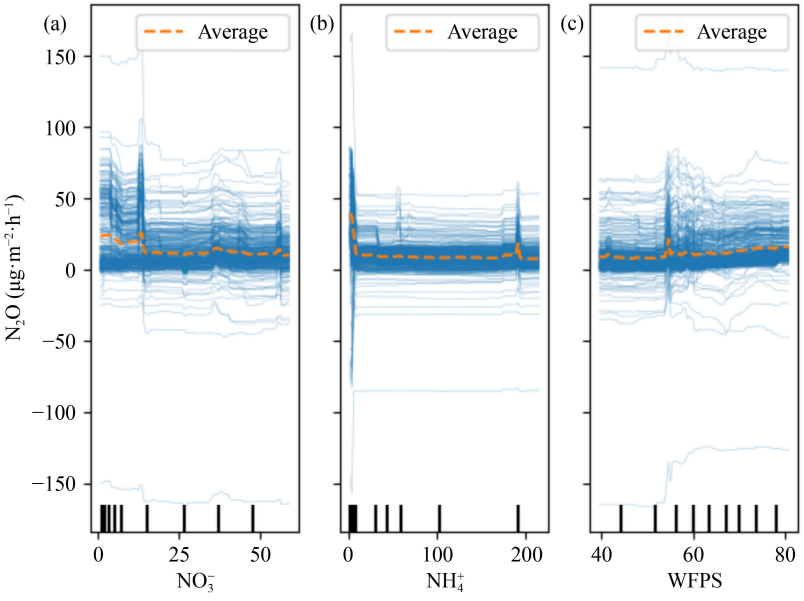


Fig. 8 Individual conditional expectation plots of predicted N_2O fluxes for (a) NO_3^-, (b) NH_4^+ and (c) water-filled pore space (WFPS).

transformation reactions are occurring, leading to the complete conversion of NH_4^+ and no relevant change in N_2O flux.

N_2O fluxes are generally higher when WFPS exceeds a threshold of 55% and ST reaches 22 °C. Our findings in this work (Figs. 8 and 9) indicate that NH_4^+ and NO_3^- have greater explanatory importance for N_2O emissions than ST and WFPS. All these four variables act as drivers of N_2O emissions, but the importance rankings generated by our model are more consistent with the view that NH_4^+ and NO_3^- are key elements involved in N_2O formation, while ST and WFPS function as environmental conditions that facilitate the biochemical transformation of NH_4^+ and NO_3^-. This finding agrees with[28], which showed that under low-oxygen conditions, as is the case in recently irrigated soils, NH_4^+ is accumulated because its transformation to NO_3^- is inhibited. Therefore, there is no increase in N_2O emission (Figs. 2 and 9).

This simulation provides essential insights into how variations in individual variables influence N_2O emissions, as captured by the model. Clearly, the proposed RF modeling of N_2O emissions from sugarcane systems has advantages over earlier modeling efforts with the same data set[4].

Colored dots represent the correlation of predictions for NH_4^+ values and the black line indicates the average trend of the data in Fig. 8. This visualization enables interpretation of the variation in the partial contribution of each characteristic. The NH_4^+ graph also serves as a reference for the color distribution of the values and facilitates visualization of the relationship between NH_4^+ concentrations and predicted N_2O emissions.

3.5 Predictability of machine learning models for assessing greenhouse gas emissions

Established models used to study N_2O flux variability are typically based on a process-level understanding of all the variable interactions and laboratory-based values for input variables. While these models are highly relevant and informative, their applicability is often limited due to their complexity and data requirements[29]. In contrast, the machine learning modeling paradigm offers a powerful set of nonlinear enhanced tools, which are faster to develop, easier to apply and well-suited for predictions, and investigation of variable interactions and simulations. Algorithms such as RF enable low-complexity analyses while producing interpretable results that can explain the behavior of the variables[20]. Machine learning models also provide opportunities for generating new insights from increasingly large and rapidly growing data sets. These models can broaden and scale existing knowledge, and

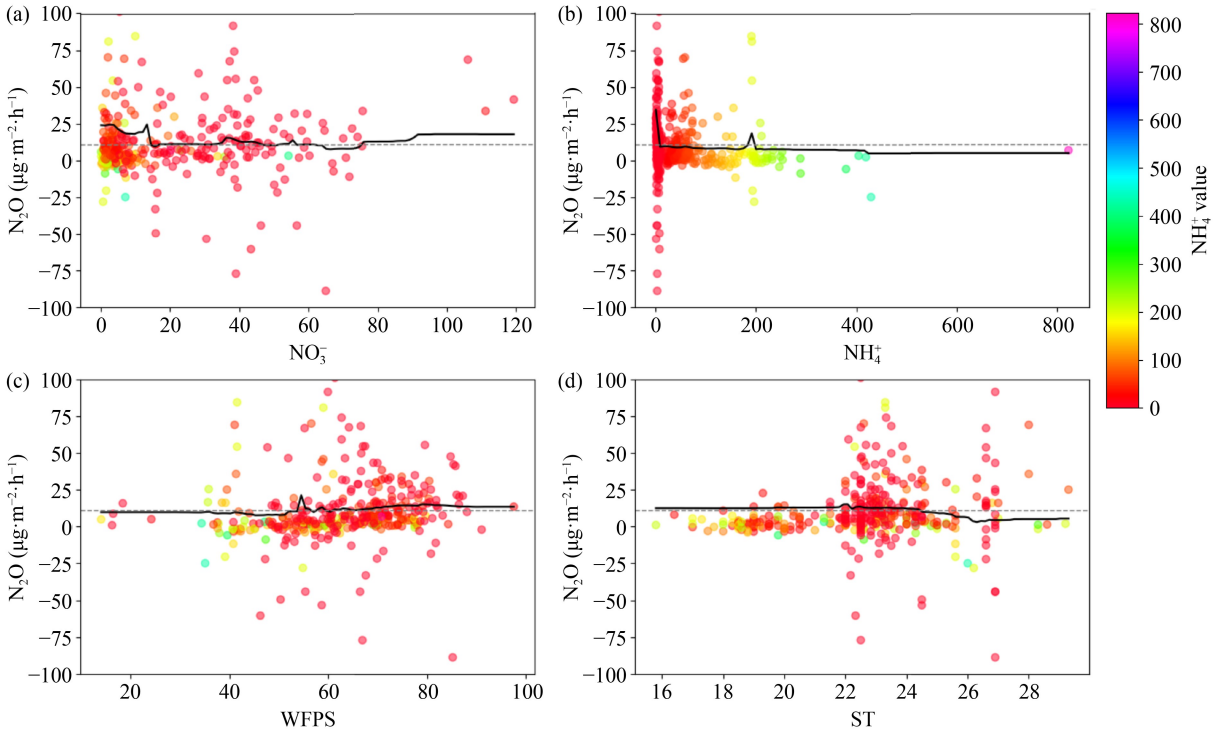


Fig. 9 Individual feature contribution extrapolations (% change in prediction) for (a) NO_3^-, (b) NH_4^+; (c) water-filled pore space (WFPS) and (d) soil temperature (ST).

have a strong potential for optimizing the study of complex variable interactions. Nevertheless, challenges remain that must be addressed to enable a broader and more effective application of machine learning techniques.

Of the variables analyzed, soil NH_4^+ availability was identified as the most influential, suggesting a strong relationship with nitrogen fertilizer application practices. This finding highlights the possibility of investigating the influence of fertilizer application strategies and subsequent nitrogen mineralization, driven by soil microbial activity, under limited soil aeration following irrigation. Also, although the relative importance of irrigation type and water volume is inferior to that of nitrogen-related variables, their influence remains noteworthy. The results showed that moderate irrigation levels do not affect emissions. However, increasing irrigation to the highest level led to a doubling of N_2O emissions, alongside a 29% increase in water expenditure with yield gains of only 17%. These observations suggested that more efficient water management strategies, for example, such as adopting the 46% irrigation treatment, could substantially reduce emissions without compromising yields.

3.6 Challenges and limitations in applying machine learning models

Data-driven machine learning models, as demonstrated in this study, are capable of nonlinear predictions of N_2O emissions. However, a key limitation is the need for larger and more diverse data sets to study N_2O emissions under a wider range of environmental conditions and over extended time periods to identify more relationships. Although this study showed that valuable insights can be gained, even with a limited set of measurements, the availability of more comprehensive data sets will be essential for further achievements.

Another prospect lies in extending the applicability of this approach to other crops and environmental conditions. However, the findings of this research are pertinent to the

extensive sugarcane production in the Brazilian Cerrado, the methods and tools used here can be further refined for a broader range of cropping systems. A challenge for future use of RF and other machine learning models in the study of N_2O and other greenhouse gas emissions is the need for broader data availability. Encouraging researchers to share their data sets openly would allow algorithms to be broadly tested and enhance the collective understanding with powerful analytical tools.

Our experimental approach was limited to specific conditions that represent sugarcane production in the Brazilian Cerrado. However, the findings highlight the potential for extending machine learning models to the analysis of other cropping systems. These models can complement process-based approaches by providing an additional layer of analysis that deepens understanding of biochemical processes and enhances, not replaces, the current set of tools for predictions.

4 Conclusions

Applying machine learning techniques to agricultural and environmental problems can enhance predictive accuracy and reveal complex, nonlinear relationships between variables, even when working with limited data sets. We have shown that RF facilitates the analysis of nonlinear interactions among N_2O predictor variables, a task that poses significant challenges for traditional statistical methods. This is particularly relevant for studying fluxes of N_2O, a potent greenhouse gas, in sugarcane fields under varying irrigation regimes within the Brazilian Cerrado biome.

Using the RF algorithm in this study on N_2O emission prediction in sugarcane systems, we identified the most influential variables in descending order as NH_4, NO_3^-, ST and WFPS. Irrigation at 75% of the crop water demand increased N_2O emissions. The relationships among the key variables soil, irrigation and climate variables were found to be nonlinear. The RF model achieved an R^2 value of 0.8736 in predicting N_2O fluxes, which indicated strong predictive performance.

Simulations of N_2O emissions or other environmental variables can be performed by systematically increasing or decreasing input parameters, making this approach a valuable analytical tool. The possibility of dealing with such relations between relevant variables, be it for retrospective analysis or future scenario prediction, is a major advantage of the enhanced machine learning modeling developed in this study.

For future research, our study results reinforce the need to integrate larger data sets that encompass biogeochemical processes, as well as climatic, crop and soil variability. The implementation of automated data collection tools across diverse regions could facilitate this effort. The most influential factors identified in this study are closely linked to elements of the nitrogen and carbon cycles in the soil. Incorporating process-based models into this machine learning framework could provide more accurate and insightful results since the technique can embed understanding of microbial interactions and soil health in the analysis.

Acknowledgements

This work was partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (Finance Code 001). The authors also gratefully acknowledge the partial support provided by EMBRAPA (Brazilian Agricultural Research Corporation) and the University of Brasília.

Compliance with ethics guidelines

Rafael T. Bonato, Lurdineide de A. B. Borges, Arminda M. Carvalho, Alexsandra D. Oliveira, Thaís R. Sousa, Maria L. G. Ramos, Walter Q. Ribeiro Junior, Robélio L. Marchão, Fernando A. M. Silva, and Díbio L. Borges declare that they have no conflicts of interest or financial conflicts to disclose. This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Portmann R W, Daniel J S, Ravishankara A R. Stratospheric ozone depletion due to nitrous oxide: influences of other gases. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 2012, **367**(1593): 1256–1264
2. Tian H Q, Xu R T, Canadell J G, Thompson R L, Winiwarter W, Suntharalingam P, Davidson E A, Ciais P, Jackson R B, Janssens-Maenhout G, Prather M J, Regnier P, Pan N Q, Pan S F, Peters G P, Shi H, Tubiello F N, Zaehle S, Zhou F, Arneth A, Battaglia G, Berthet S, Bopp L, Bouwman A F, Buitenhuis E T, Chang J F, Chipperfield M P, Dangal S R S, Dlugokencky E, Elkins J W, Eyre B D, Fu B J, Hall B, Ito A, Joos F, Krummel P B, Landolfi A, Laruelle G G, Lauerwald R, Li W, Lienert S, Maavara T, MacLeod M, Millet D B, Olin S, Patra P K, Prinn R G, Raymond P A, Ruiz D J, van der Werf G R, Vuichard N, Wang J J, Weiss R F, Wells K C, Wilson C, Yang J, Yao Y Z. A comprehensive quantification of global nitrous oxide sources and sinks. *Nature*, 2020, **586**(7828): 248–256
3. Butterbach-Bahl K, Dannenmann M. Denitrification and associated soil N_2O emissions due to agricultural activities in a changing climate. *Current Opinion in Environmental Sustainability*, 2011, **3**(5): 389–395
4. de Carvalho A M, de Oliveira A D, Coser T R, de Sousa T R, de Lima C A, Ramos M L G, Malaquias J V, de Araujo Gonçalves A D M, Ribeiro W Q Jr. N_2O emissions from sugarcane fields under contrasting watering regimes in the Brazilian savannah. *Environmental Technology & Innovation*, 2021, **22**: 101470
5. Yin Y L, Wang Z H, Tian X S, Wang Y C, Cong J H, Cui Z L. Evaluation of variation in background nitrous oxide emissions: a new global synthesis integrating the impacts of climate, soil, and management conditions. *Global Change Biology*, 2022, **28**(2): 480–492
6. Organisation for Economic Co-operation and Development (OECD), Food and Agriculture Organization of the United Nations. OECD-FAO Agricultural Outlook 2023–2032. Paris: *OECD Publishing*, 2023
7. Friedl J, Warner D, Wang W J, Rowlings D W, Grace P R, Scheer C. Strategies for mitigating N_2O and N_2 emissions from an intensive sugarcane cropping system. *Nutrient Cycling in Agroecosystems*, 2023, **125**(2): 295–308
8. IPCC Guidelines for National Greenhouse Gas Inventories (IPCC). Prepared by the National Greenhouse Gas Inventories Programme. In: Eggleston H S, Buendia L, Miwa K, Ngara T, Tanabe K, eds. Published: *IGES*, 2006
9. Bouwman A F, Boumans L J M, Batjes N H. Modeling global annual N_2O and NO emissions from fertilized fields. *Global Biogeochemical Cycles*, 2002, **16**(4): 28-1–28-9
10. Wang C, Amon B, Schulz K, Mehdi B. Factors that influence nitrous oxide emissions from agricultural soils as well as their representation in simulation models: a review. *Agronomy*, 2021, **11**(4): 770
11. Parton W J, Holland E A, Del Grosso S J, Hartman M D, Martin R E, Mosier A R, Ojima D S, Schimel D S. Generalized model for NO_x and N_2O emissions from soils. *Journal of Geophysical Research*, 2001, **106**(D15): 17403–17419
12. Li Y, Chen D L, Zhang Y M, Edis R, Ding H. Comparison of three modeling approaches for simulating denitrification and nitrous oxide emissions from loam-textured arable soils. *Global Biogeochemical Cycles*, 2005, **19**(3): 1–15
13. Arnold J G, Srinivasan R, Muttiah R S, Williams J R. Large area hydrologic modeling and assessment Part I: model development. *Journal of the American Water Resources Association*, 1998, **34**(1): 73–89
14. Pattey E, Edwards G C, Desjardins R L, Pennock D J, Smith W, Grant B, MacPherson J I. Tools for quantifying N_2O emissions from agroecosystems. *Agricultural and Forest Meteorology*, 2007, **142**(2–4): 103–119
15. Scarpare F V, Hernandes T A D, Ruiz-Corrêa S T, Picoli M C A, Scanlon B R, Chagas M F, Duft D G, de Fátima Cardoso T. Sugarcane land use and water resources assessment in the expansion area in Brazil. *Journal of Cleaner Production*, 2016, **133**: 1318–1327
16. Fonseca da Silva J, Moreira de Carvalho A, Rein T A, Rodrigues Coser T, Quadros Ribeiro W J, Lino Vieira D, Coomes D A. Nitrous oxide emissions from sugarcane fields in the Brazilian Cerrado. *Agriculture, Ecosystems & Environment*, 2017, **246**: 55–6517
17. Saha D, Basso B, Robertson G P. Machine learning improves predictions of agricultural nitrous oxide (N_2O) emissions from intensively managed cropping systems. *Environmental Research Letters*, 2021, **16**(2): 024004
18. Lawrence N C, Tenesaca C G, VanLoocke A, Hall S J. Nitrous oxide emissions from agricultural soils challenge climate sustainability in the US corn belt. *Proceedings of the National Academy of Sciences of the United States of America*, 2021, **118**(46): e2112108118
19. Abalos D, Rittl T F, Recous S, Thiébeau P, Topp C F E, van Groenigen K J, Butterbach-Bahl K, Thorman R E, Smith K E, Ahuja I, Olesen J E, Bleken M A, Rees R M, Hansen S. Predicting field N_2O emissions from crop residues based on their biochemical composition: a meta-analytical approach. *Science of the Total Environment*, 2022, **812**: 152532
20. Breiman L. Random forests. *Machine Learning*, 2001, **45**(1): 5–32
21. Fischer A. How to determine the unique contributions of input-variables to the nonlinear regression function of a multilayer perceptron. *Ecological Modelling*, 2015, **309–310**:

60–63
22. Belgiu M, Drăguţ L. Random forest in remote sensing: a review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2016, **114**: 24–31
23. Towfiqul Islam A R M, Talukdar S, Mahato S, Kundu S, Eibek K U, Pham Q B, Kuriqi A, Linh N T T. Flood susceptibility modelling using advanced ensemble machine learning models. *Geoscience Frontiers*, 2021, **12**(3): 101075
24. Fan G F, Zhang L Z, Yu M, Hong W C, Dong S Q. Applications of random forest in multivariable response surface for short-term load forecasting. *International Journal of Electrical Power & Energy Systems*, 2022, **139**: 108073
25. Alpaydın E. *Introduction to Machine Learning*. Cambridge, Mass: *MIT Press*, 2004
26. Github. Prediction-of-Sugarcane-N_2O-Emissions. USA, 2025. Available at Github website on February 20, 2025
27. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 2015, **24**(1): 44–65
28. Firestone M, Davidson E. Microbiological basis of NO and N_2O production and consumption in soil. In: Andreae M, Schinel D, eds. Exchange of Trace Gases Between Terrestrial Ecosystems and the Atmosphere. New York: *John Wiley & Sons*, 1989, 7–21
29. Gaillard R K, Jones C D, Ingraham P, Collier S, Izaurralde R C, Jokela W, Osterholz W, Salas W, Vadas P, Ruark M D. Underestimation of N_2O emissions in a comparison of the DayCent, DNDC, and EPIC models. *Ecological Applications*, 2018, **28**(3): 694–708