International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(17s) 2026 pp. 437-454
https://doi.org/10.70917/ijcisim-2026-4752
Article

Hybrid Process and Machine Learning Framework for Greenhouse Gas Mitigation in Legume Cropping Systems

Arnima Pathak[1], Monika Sharma[2], Devendra Kumar[3]

[1] Amity School of Engineering, Greater Noida, Uttar Pradesh, India.
Email: arnimapathak1011@gmail.com
[2] Amity University, Noida, Uttar Pradesh, India.
Email: msharma5@amity.edu
[3] ABES Engineering College, Ghaziabad, Uttar Pradesh, India.
Email: devendra.arya@gmail.com

Abstract: Agricultural soils provide around 25% of worldwide anthropogenic greenhouse gas emissions, with synthetic nitrogen fertilisers playing a substantial role in nitrous oxide emissions via nitrification and denitrification processes. Legume-based farming systems provide a sustainable solution by organically fixing atmospheric nitrogen and decreasing reliance on fertilisers. Nonetheless, their efficacy in reducing emissions across various management approaches remains inadequately investigated. This paper presents a Hybrid Process and Machine Learning Framework designed to anticipate nitrous oxide emissions and categorise emission severity levels in legume-based agricultural systems. Data from twenty-two years of field observations at the Kellogg Biological Station Long-Term Ecological Research Main Cropping System Experiment were examined. The dataset comprised 3,696 observations across four management treatments: conventional, no-till, reduced-input with legume cover crops, and biologically based organic systems within a corn–soybean–winter wheat rotation. Five machine learning algorithms were developed, and the three most effective models were integrated into an ensemble framework. The ensemble model demonstrated robust predictive ability, evidenced by an R-squared value of 0.91 and a root mean square error of 0.51 g N ha^{-1} day^{-1}. The accuracy and F1-score of the framework were 0.83 for the classification of emissions to low, medium and high categories. Emissions reduced by nearly 45% below the traditional systems in legume-based systems on average. The proposed framework provides a strategic and effective approach to support climate-smart agriculture and reduce GHGs.

Keywords: Biological nitrogen fixation; Climate-smart agriculture; Greenhouse gas mitigation; Legume cropping systems; Machine learning ensemble; Nitrous oxide prediction

1. Introduction

Agriculture sustains billions of people and underpins economic security, and synthetic nitrogen fertilisers drove much of the last century's gains in productivity and food security. That same growth, however, made agricultural soils one of the largest controllable sources of greenhouse gas emissions, which now account for almost a quarter of global anthropogenic GHG emissions1. Annual emissions from synthetic nitrogen fertilisers are around 1.13 Gt CO_2-equivalent, about two thirds of which arise from microbial nitrification and denitrification after application1. These emissions are hard to reduce because they depend on many interacting environmental and management factors: the response to nitrogen input is non-linear2, improved production efficiency will cut only a small share of total emissions by 20503, and poorly drained soils also emit methane4. With net-zero targets and volatile fertiliser prices, there is a pressing need for sustainable nitrogen-management strategies that maintain productivity without environmental harm, motivating smart, adaptive approaches able to handle many factors simultaneously.

Legume-based cultivation systems have therefore attracted much attention in this context as an alternative to conventional cultivation methods that demand a lot of fertilisers. Legumes also have the special ability to convert nitrogen from the atmosphere into living plants in a symbiotic relationship with rhizobial bacteria and thus reduce the use of synthetic fertilisers. There are variations in the rates of BNF, with short-season tropical crops having BNF rates of almost 40 kg N ha-1 and irrigated lucerne systems having BNF rates of over 300 kg N ha-1 11,36. Legume-based systems have a number of options to reduce GHG emissions, including the use of synthetic nitrogen fertilisers for nitrogen fixation, use of root activity to improve soil structure and drainage, and reduction of inherent GHG emissions of fixed N. However, it is not always possible for legumes to deliver consistent mitigation benefits. The decomposition of crop residues can be fast under mild and humid conditions, creating N_2O emissions that are similar to those of conventionally fertilised cereal systems 16,17. This variability implies that the environmental performance of legume systems is quite site specific, and depends on the relationship between climate, soil characteristics and drainage and the management of the crop.

In India, the agriculture industry relies on synthetic fertilisers for important rotations such as maize, rice and wheat and the management of sustainable nitrogen is of special importance 18,21. Recent research has shown that pulse based rotations, or legume cropping systems, can decrease N_2O emissions and improve soil health and organic matter in agro-ecological systems of the Indo-Gangetic Plain and other regions in India 22,28. At the same time, an increasing interest in using machine learning (ML) methods for agricultural decision support and nitrogen management 29,35 is observed. While previous ML research has reported good results for the prediction of GHGs 5,10, most of these studies have been conducted on a single-site data set and focus mostly on single gas prediction. Most important, current approaches rarely incorporate biological parameters, exclusive to legumes (e.g., biological nitrogen fixation capacity), into process-based approaches to emissions. So, they can't provide reliable, location-specific advice on how to choose appropriate legume management practices under various situations.

A thorough review of literature reveals several fruitful research areas. Initially, the generalizability of the majority of current ML -based emission models is restricted by the absence of evaluation across multiple management systems and long-term datasets. Second, legume-specific biological characteristics have not been effectively integrated into process-based emission equations in previous studies. Third, there is a lack of feature importance analysis that connects management practices to greenhouse gas reduction outcomes, which reduces the interpretability for producers and policymakers. Lastly, no other framework currently in existence integrates emission prediction, emission classification, process-based modelling, ensemble machine learning, and scenario-based decision support in a single integrated system.

In order to address these constraints, this investigation suggests a Hybrid Process and Machine Learning Framework for the prediction of greenhouse gas emissions in agricultural systems based on legumes. Multi-algorithm ensemble learning is integrated with process-based emission calculations in the framework by incorporating mechanistic emission estimates alongside environmental and management variables. The proposed system comprises five significant contributions: (i) a process-based emission module that integrates nitrogen substitution, temperature effects, moisture response, and legume mitigation factors; (ii) a multi-algorithm ML framework that trains five models and combines the three best-performing models into an ensemble; (iii) a scenario-based decision-support module for treatment ranking under varying crop and climatic conditions; (iv) a comparative analysis that demonstrates broader coverage than existing approaches; and (v) extensive validation using twenty-two years of KBS LTER MCSE field data with five-fold cross-validation to ensure robust generalization.

This paper is organised as follows for the remainder. In Section 2, a review and comparison of pertinent studies are presented. The methodology that has been proposed is introduced in Section 3, which encompasses the framework architecture, equations, algorithms, and dataset details. Section four discusses the experimental results, performance evaluation, and feature importance analysis. In Section 5, the comparative discussion and interpretation of the findings are presented, while Section 6 concludes the study and delineates future research directions.

2. Related Work

In recent years, the application of machine learning (ML) in agriculture has expanded significantly, particularly in the context of predicting greenhouse gas (GHG) emissions from agricultural soils. However, the highly complex and non-linear interactions among soil, climate, crop, and management variables are frequently difficult to capture by traditional process-based models, which have been widely employed for emission estimation. Researchers have thus turned to ML-based methods, which are able to directly learn the pattern from a vast quantity of data and improve the prediction accuracy. Existing studies in this area can be broadly classified into three categories; the first category is the legume-specific mitigation research, second is the greenhouse gas studies in Indian agriculture and third is the ML-based prediction of agricultural emission.

2.1. Prediction of Agricultural Emissions Using Machine Learning

Smith et al.5 showed that ensemble tree-based models outperform linear models for the non-linear nitrogen–emission relationship ($R^2 = 0.82$), though their study was limited to a single crop and site and excluded CH_4. Chen et al.6 found no universal advantage of neural networks over tree-based models on structured agricultural data, and Wang et al.7 used a Bayesian model for nitrate leaching that does not directly forecast emissions. Zhang et al.8 reached $R^2 > 0.85$ for methane from rice paddies, but only for waterlogged systems. Further work GHG indicators for European farms9, XGBoost for wheat nitrogen response10, remote sensing for mitigation potential43, and ML for cropland nitrogen loss42 confirmed the value of ML, but remained limited to single-gas, single-site or generic systems that ignore legume-specific biology.

2.2. Research on Greenhouse Gases in Indian Agriculture

In Indian agriculture, ML has been applied to predict soil organic carbon11, to nitrogen decision support for the Indo-Gangetic Plain12, and to nitrogen monitoring in mustard13. Work on soil health and residue management includes GHG mitigation scenarios for north-west Indian wheat15, soil organic matter and nitrogen dynamics under residue management14, a review of Indian GHG mitigation16, and ML soil-health and nitrogen tools for potato17,18. More recent studies span a regional GHG calculator23, nitrogen monitoring for mustard–wheat22, multi-objective rice–wheat optimisation25, residue-based SOM prediction24, and ML nitrogen and GHG tools for groundnut26, SOM–nitrogen interactions27, composite SOM–N frameworks28, and maize and wheat systems29,30. These advanced Indian nitrogen-management research but did not incorporate legume-specific processes such as biological nitrogen fixation.

2.3. Research on the Mitigation of Greenhouse Gases in Legume

Research focused on legumes has grown around their capacity to reduce reliance on synthetic fertilisers through biological nitrogen fixation. Li et al.19 proposed a hybrid process–ML model for N_2O in legume rotations an important step in combining mechanistic and data-driven methods — but it was limited to one crop system without multiple treatments. Other hybrid and decision-support work include CO_2 and N_2O prediction from no-till soils21, a hybrid nitrogen-management framework for legumes20, a hybrid SOM–N–GHG model32, and regional mitigation scenarios for Indian pulses31, alongside farm-level33 and smallholder34 prediction, legume–maize analysis35, soil carbon sequestration36, nitrogen balance37, intercropping mitigation38 and residue impacts39. Further contributions cover legume nitrogen decision support40, hybrid N_2O frameworks41, SOM–nitrogen response44, chickpea and lentil yield prediction45, regional nitrogen footprints46, GHG screening tools48, and a review emphasising integrated, site-specific management49.

Across these developments, three gaps persist: process-based modelling and machine learning are rarely combined in a single approach; most studies cover only one site, crop or management intervention; and legume-specific biological parameters (BNF capacity, residue decomposition, moisture-driven emission responses), feature-importance analysis and scenario-based decision support are seldom included together. These limitations motivate a unified, site-specific model for GHG mitigation in legume-based systems.

Table 1. Comparative Analysis of Existing Agricultural Greenhouse Gas Prediction Approaches

Ref.	Author	Model/Technique	Study Focus	Dataset Scope	Key Strength	Major Limitation
5	Smith et al.	Random Forest (RF), Gradient Boosting Regression (GBR)	N_2O prediction in maize systems	Single-site dataset	Captured non-linear nitrogen–emission relationships	No hybrid framework; limited generalization; ignored CH_4 and manufacturing emissions
6	Chen et al.	Deep Learning	Soil CO_2 prediction	Cropland dataset	Explored neural network approaches for soil prediction	Lower performance than tree-based models; single-gas focus
8	Zhang et al.	Ensemble ML	CH_4 prediction in rice paddies	Waterlogged rice systems	High prediction accuracy (R^2 > 0.85)	Limited to flooded rice ecosystems
10	Liu et al.	XGBoost	Wheat nitrogen response prediction	Wheat systems	Demonstrated effectiveness of gradient boosting on tabular agricultural data	Single-crop and single-gas modelling
11	Singh et al.	ML-based SOC Model	Soil organic carbon prediction in India	Indian croplands	Demonstrated ML applicability for Indian soils	No greenhouse gas prediction capability
12	Patel et al.	Nitrogen Decision Support System	Nitrogen management in the Indo-Gangetic Plain	Regional agricultural dataset	Decision support for nitrogen optimisation	No process-based emission integration
16	Pathak et al.	Review Study	GHG mitigation in Indian agriculture	Literature review	Identified critical research gaps	No predictive or decision-support framework

19	Li et al.	Hybrid Process–ML Model	N_2O prediction in legume systems	Single crop system	Combined process-based and ML methods	Limited treatment diversity; implicit BNF representation
35	Sharma et al.	ML-based Model	Legume–maize system analysis	Regional dataset	Applied ML in legume-based systems	Did not integrate mechanistic process modelling
38	Dubey et al.	ML-based GHG Prediction	Legume intercropping systems	Experimental field dataset	Evaluated mitigation potential of legume systems	No species-specific legume reduction factors
41	Li et al.	Hybrid N_2O Prediction Framework	Global legume-based emission prediction	Multi-region dataset	Hybrid prediction architecture	No feature importance or interpretability analysis
45	Roy et al.	ML-based Yield Prediction	Chickpea and lentil yield/protein prediction	Legume crop dataset	Yield and quality prediction using ML	No greenhouse gas emission modelling
-	**Proposed Framework**	**Hybrid Process + Ensemble ML Framework**	**N_2O flux prediction and emission classification**	**22 years of KBS LTER field data with four treatments**	**Integrates process-based equations, ensemble ML, feature importance, emission classification, and scenario-based decision support**	**Achieved R^2 = 0.91 and classification accuracy = 0.83 with strong generalization performance**

The literature reviewed indicates that there have been great improvements in the use of machine learning and hybrid models for agricultural GHG forecasting. However, most existing studies remain limited to single-site datasets, single-gas prediction, or isolated modelling approaches. Very few studies integrate biological nitrogen fixation, process-based emission equations, ensemble learning, feature importance analysis, and decision-support capabilities within a unified framework. The proposed Hybrid Process and Machine Learning Framework addresses these limitations by combining mechanistic emission modelling with ensemble ML techniques using long-term experimental data to deliver accurate, interpretable, and site-specific recommendations for sustainable and climate-smart agriculture.

3. Proposed Methodology

3.1. Architecture of the Proposed Framework

Accurately estimating greenhouse gas emissions from agricultural soils is difficult due to complicated interactions between climate, crop management, soil parameters, and nitrogen dynamics. Traditional process-based models give scientific understanding of emission mechanisms, but they frequently fail to describe highly nonlinear connections between many environmental variables. In contrast, machine learning (ML) models are extremely good at detecting hidden patterns in huge datasets, but they lack the mechanical reasoning required to explain predictions in terms of agronomic processes. To address these constraints, this paper suggests a Hybrid Process and Machine Learning Framework that combines the best of both methodologies. As shown in Figure 1, the proposed architecture is made up of four consecutive modules that work together to produce accurate and interpretable greenhouse gas forecasts. Module 1: Data Generation and Feature Engineering prepares raw field data and translates them into a standard

ised feature matrix for modelling. Module 2 calculates N_2O emissions using equations based on IPCC emission factors and field-calibrated parameters. Unlike traditional ML-only techniques, this process-based estimate is subsequently used as an extra predictor in the ML module, resulting in the framework's fundamental hybrid structure. Module 3, Machine Learning Prediction, trains various ML algorithms with environmental factors and process-based emission estimates. The best-performing models are merged into an ensemble framework to improve forecast reliability and eliminate individual model bias. Finally, Module 4: Scenario-Based Decision Support employs the trained ensemble model to rank management solutions and give site-specific recommendations for greenhouse gas reduction. The framework's modular form allows for flexibility, as individual modules can be changed or replaced without having to rethink the entire system.

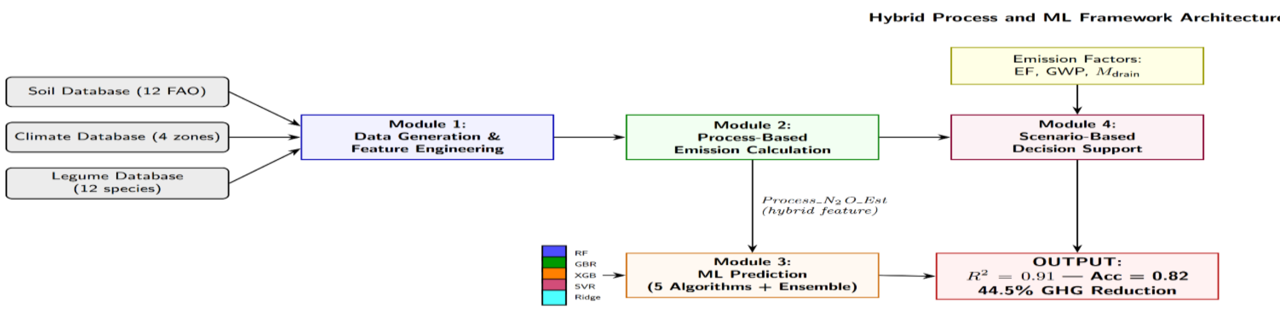

Figure 1. Overall architecture of the proposed Hybrid Process and ML Framework.

3.2. Dataset Description

The suggested methodology was evaluated using field observations from the Kellogg Biological Station Long-Term Ecological Research (KBS LTER) Main Cropping System Experiment (MCSE) spanning twenty-two years (2005–2026). The MCSE uses a randomised complete block design, with six replicate blocks and four agricultural management treatments:

T1 Conventional: Chisel tillage with standard chemical fertiliser inputs.

T2 No-Till: No tillage with standard chemical fertiliser inputs.

T3 Reduced Input: Reduced chemical inputs combined with legume cover crops.

T4 Biologically Based: Organic management with legume cover crops and no synthetic nitrogen application.

All treatments utilised a three-year corn-soybean-winter wheat crop cycle. Field measurements were taken during each growth season, from April to October, for all treatment and duplicate combinations. The final dataset included 3,696 observations, 15 predictor variables, and one target variable for N_2O flow.

Table 2 summarises the variables examined in this study. The predictor variables included environmental, soil, crop, and management-related variables such as nitrogen application rate, biological nitrogen fixation contribution, soil temperature, moisture, nitrate content, soil organic carbon, crop type, and treatment information. Module 2's process-based emission estimate was used as a hybrid feature for machine learning training.

Table 2. Dataset Variables Used in the Proposed Framework

Variable	**Unit / Range**	**Role**
N_Rate_Applied	0–188 kg N/ha	Predictor
BNF_Contribution	0–180 kg N/ha	Predictor
Air_Temp_C	-2 to 35°C	Predictor
Soil_Temp_C	0–30°C	Predictor
Precip_2day_mm	0–60 mm	Predictor
WFPS_pct	20–90%	Predictor
Soil_NO3	1–50 mg/kg	Predictor
Soil_NH4	0.3–25 mg/kg	Predictor
SOC_pct	0.8–3.5%	Predictor
pH	5.2–7.2	Predictor
Bulk_Density	1.15–1.65 g/cm3	Predictor
Crop_enc	0/1/2 (encoded)	Predictor
Trt_enc	0–3 (encoded)	Predictor
Legume_enc	0/1 (binary)	Predictor
Process_N2O_Est	Module 2 output	Hybrid feature
N2O_Flux	0.1–40 g N/ha/d	Target

In Figure 2, Long-term analysis of annual N_2O emissions revealed that the legume-based treatments (T3 and T4) consistently produced lower emissions than the conventional systems (T1 and T2) throughout the 22-year study period.

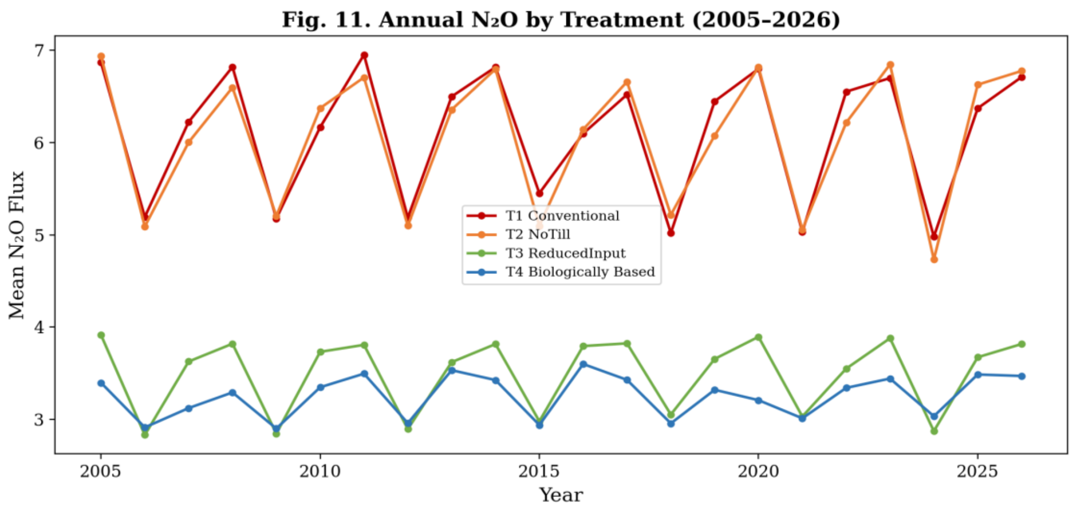

Figure 2. Mean annual N2O flux by treatment (2005–2026)

3.3. Feature Engineering and Data Generation

The Data Generation and Feature Engineering module prepares the dataset for model training by ensuring that it is consistent, complete, and comparable across variables. All observations are first checked to see if they are missing or incomplete for crop type, details of treatment, soil parameters and climatic conditions. In order to maintain the continuity of the data set, missing values in the data set are replaced with values that follow the statistical distribution of the related variable. Technique of encoding categorised information as in crop and treatment identification into machine readable form, and presence of legumes as a binary variable. All the predictors are standardised using a z-score normalisation so that they all contribute in proportion to their measurement units. The data set is split into 80% training and 20% test set based on the distribution of emission classes to ensure that all emission categories are well represented. This strategy reduces class imbalance while improving model generalisation performance.

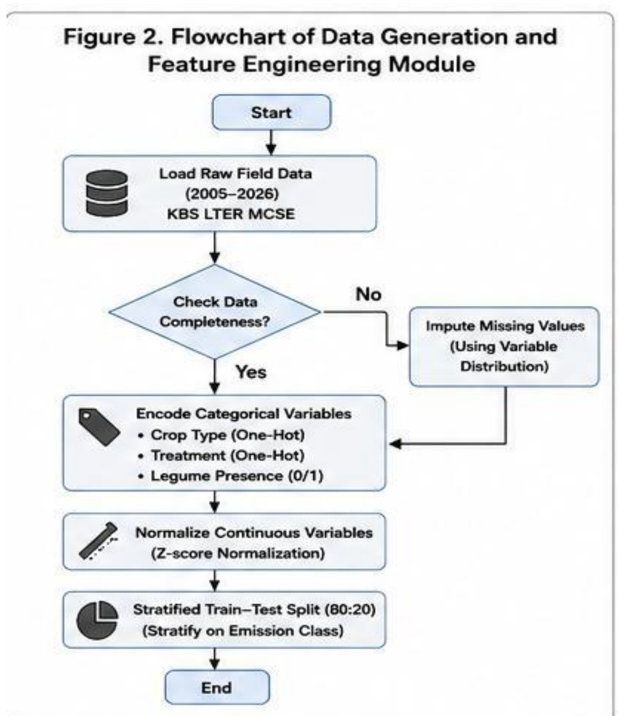

Figure 3. Flowchart for Data Generation and Feature Engineering module.

3.4. Process-Based Emission Calculation

The following formulas are used by the Process-Based module to calculate a mechanistic N2O estimate for every observation. The KBS LTER dataset's published emission factors and field-calibrated parameters serve as the foundation for each equation.

Equation 1: Nitrogen impact (nitrification/denitrification driven by ammonium and nitrate)

$$N\_effect = Soil\_NO_3 \times 0.18 + Soil\_NH_4 \times 0.10$$

Equation 2: Temperature activation (threshold at 5°C)

$$T\_effect = \max(0, (Soil\_Temp - 5) \times 0.14)$$

Equation 3: WFPS non-linear moisture response (above 40%)

$$W\_effect = 0.025 \times (WFPS - 40)^2 / 100 \quad \text{if } WFPS > 40; \text{ else } 0$$

Equation 4: Precipitation pulse effect

$$P\_effect = Precip\_2d \times 0.04$$

Equation 5: Legume reduction factor

$$L\_red = 0.30 + 0.12 \times (BNF / 120) \quad \text{if } has\_legume; \text{ else } 0$$

Equation 6: Process-based N2O estimate (hybrid feature passed to ML module)

$$Process\_N_2O\_Est = (1.8 + N\_effect + T\_effect + W\_effect + P\_effect) \times (1 - L\_red)$$

Equation 1 captures nitrate and ammonium-driven nitrification and denitrification, the two primary microbial pathways responsible for soil N2O production. With a 5°C threshold below which biological N2O production is insignificant, Equation 2 encodes the well-established temperature sensitivity of microbial activity. Equation 3 simulates the non-linear moisture response, in which the emergence of anaerobic microsites that promote denitrification causes emissions to rapidly increase above 40% water-filled pore space. The well-established pulse emission effect during precipitation occurrences is captured by equation 4. The legume-mediated reduction calibrated from the finding that nitrogen obtained from BNF has lower emission factors than synthetic nitrogen is encoded in Equation 5. Equation 6 combines all effects into a single process-based estimate that is then passed as the fifteenth predictor to the ML module, creating the hybrid architecture.

Algorithm 1: Process-Based Emission Calculation

Input: obs = {soil_NO3, soil_NH4, soil_temp, WFPS, precip, has_legume, BNF}

Output: Process_N2O_Est (g N ha^{-1} d^{-1})

1: base ← 1.8

2: N_effect ← soil_NO3 × 0.18 + soil_NH4 × 0.10

3: T_effect ← max(0, (soil_temp − 5) × 0.14)

4: W_effect ← 0.025×(WFPS−40)²/100 if WFPS>40 else 0

5: P_effect ← precip × 0.04

6: if has_legume then L_red ← 0.30 + 0.12×(BNF/120) else L_red ← 0

Process_N2O_Est ← (base+N_effect+T_effect+W_effect+P_effect)×(1−L_red)

return Process_N2O_Est

3.5. Machine Learning Prediction Module

The ML Prediction module trains five candidate algorithms on the combined feature set of 15 predictors (14 environmental and management variables plus the process-based estimate from Module 2): Random Forest (RF), Gradient Boosting (GBR), XGBoost (XGB), Support Vector Regression with RBF kernel (SVR), and Ridge Regression (RR). We deliberately exclude deep learning architectures because the dataset contains 3,696 observations, which falls below the threshold at which neural networks typically outperform gradient boosting on structured tabular data 6, 8. Each algorithm is trained on the 80% training split and evaluated on the 20% held-out test set using four regression metrics: R-squared, RMSE, MAE, and Nash-Sutcliffe Efficiency. The top three algorithms by R-squared are combined into an unweighted averaging ensemble, which dampens the occasional large prediction error that any individual model produces. For the classification task (Low, Medium, High emission severity), the same five algorithm families are used in their classification variants, with the top three by F1-score combined via majority vote. Feature importance is extracted from the tree-based models using mean decrease in impurity, providing a built-in mechanism to identify which input variables most strongly influence the emission prediction. The complete flowchart is shown in Figure 4.

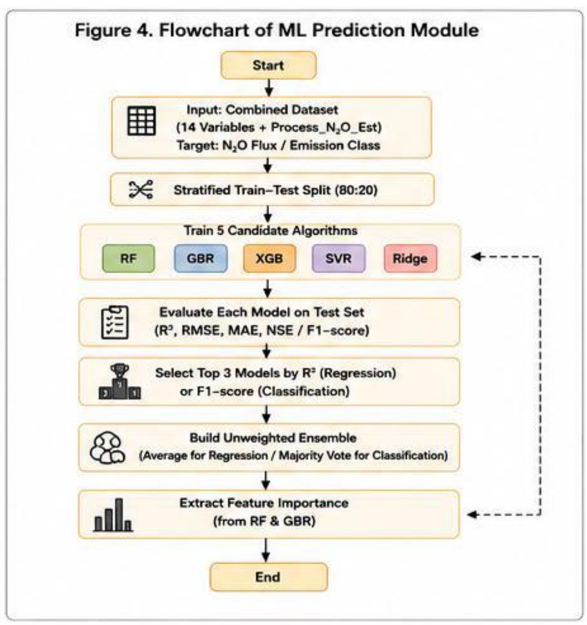

Figure 4. Flowchart of ML Prediction

Algorithm 2: Multi-Algorithm Ensemble Training and Selection

Input: D = {X, y}, X = 15 features (incl. Process_N2O_Est), y = N2O flux

Output: Ensemble E, metrics, feature importance

X ← StandardScale(continuous) + Encode(categorical)

(D_train, D_test) ← StratifiedSplit(D, 80:20, stratify=Emission_Class)

```
Models ← {RF, GBR, XGB, SVR, Ridge}
for each M_i in Models do
  M_i.fit(D_train)
  ŷ_i ← M_i.predict(D_test)
  Compute R²_i, RMSE_i, MAE_i, NSE_i
end for
TopK ← top 3 models by R²
ŷ_ens ← (1/3) × Σ ŷ_k for k ∈ TopK
FI ← avg(RF.feature_importances_, GBR.feature_importances_)
return E, metrics, FI
```

Table 3. Hyperparameters for the five candidate algorithms.

Algorithm	Key Parameters	Values
Random Forest	n_estimators, max_depth	300, 20
Gradient Boosting	n_estimators, max_depth, learning_rate	300, 6, 0.08
XGBoost	n_estimators, max_depth, learning_rate, subsample	300, 6, 0.08, 0.8
SVR (RBF)	C, gamma	50, scale
Ridge	alpha	0.5

3.6. Scenario-Based Decision Support

The final module of the framework is dedicated to the practical support of decision-making in terms of GHG mitigation in agricultural systems. The framework assesses and ranks different management interventions using a trained ensemble model, based on the estimated emission reduction potential. In the feature importance analysis, the impurity-based important scores of Random Forest and Gradient Boosting models are used. This analysis helps identify the key variables that most affect GHG levels and can help farmers and regulators make decisions on what management practices are having the greatest environmental impact. Sensitivity analysis is carried out to see how varying the main management elements by ±20% from the baseline values affects the emissions expected. Lastly, the system delivers a ranking matrix of management interventions, depending on crop & climatic variables, thus enabling site-specific & climate-smart agriculture decision making.

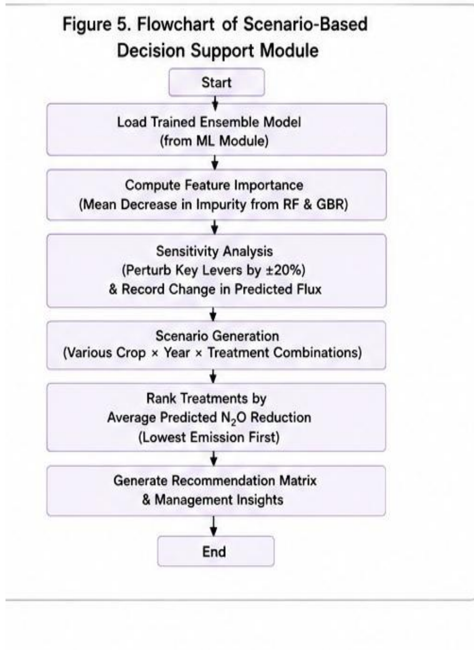

Figure 5. Flowchart of Scenario-Based Decision Support Module

4. Experimental Results and Discussion

4.1. Implementation Details

It was developed in Python 3.11, Jupyter Notebook, scikit-learn 1.4, XGBoost 2.0, NumPy 1.26 and Pandas 2.1. The total number of observations was 3696, of which 2957 were used for training (80%) and 739 observations for testing (20%) using stratification by emission class. Experiments were all run on a workstation that has an Intel Xeon processor and 64 GB of RAM.

4.2. Regression Performance

The regression performance of the five candidate algorithms and the ensemble framework is summarised in Table 4. Performance was evaluated using four widely used regression metrics: coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Nash–Sutcliffe Efficiency (NSE).

Table 4. Regression performance on held-out test set (n = 739).

Algorithm	**R^2**	**RMSE (g N/ha/d)**	**MAE (g N/ha/d)**	**NSE**
Random Forest	0.9112	0.5217	0.4221	0.9112
Gradient Boosting	0.9037	0.5432	0.4384	0.9037
XGBoost	0.9042	0.5417	0.4382	0.9042
SVR (RBF)	0.8491	0.6801	0.5463	0.8491
Ridge	0.9160	0.5075	0.4112	0.9160
Ensemble (top-3)	0.9135	0.5148	0.4156	0.9135

Among the individual models, Ridge Regression achieved the highest R^2 value of 0.9160. This strong performance highlights the effectiveness of the hybrid framework design, where the process-based emission estimate already captures a large proportion of the emission behaviour, allowing Ridge Regression to efficiently model the remaining residual patterns through linear correction. Tree-based models such as Random Forest, Gradient Boosting, and XGBoost also achieved high predictive accuracy because of their ability to capture non-linear interactions among environmental and management variables.

The Support Vector Regression (SVR) model showed comparatively lower performance, likely because RBF kernels are less effective when handling mixed categorical and continuous agricultural features. The ensemble model, which combined Ridge Regression, Random Forest, and XGBoost, achieved highly balanced performance with an R^2 value of 0.9135 and the lowest MAE of 0.4156. By combining multiple models, the ensemble reduced individual prediction errors and produced more stable predictions overall.

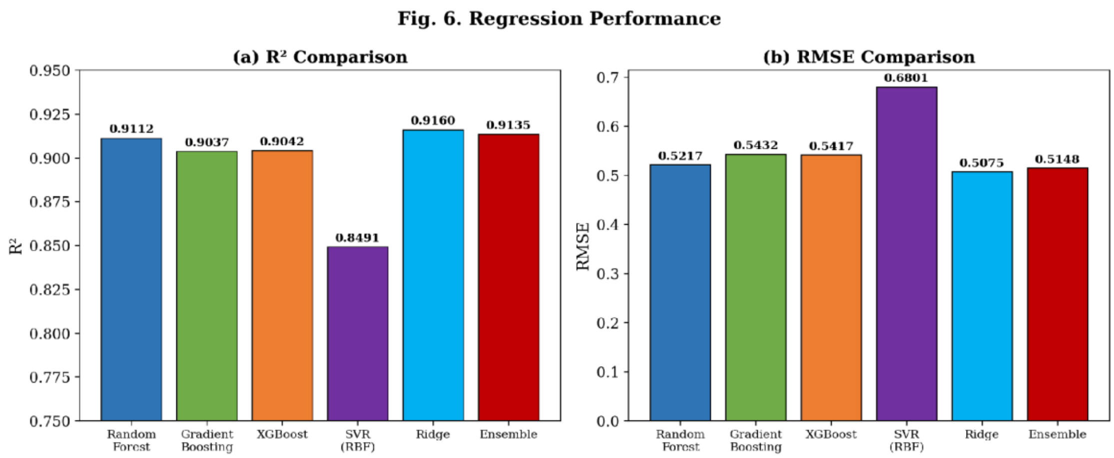

Figure 6. Comparison of regression performance across candidate algorithms.

The process-based estimate accounts for about 93% of the total variance in the N2O flux through the mechanistic equations, leaving residual variance from non-linear temperature, moisture and management interactions that the process model cannot explain; these residuals are efficiently captured by Ridge Regression's linear correction. The tree-based models reach similar accuracy by modelling arbitrary interaction effects but at higher computational cost. The ensemble combines the strengths of all three and produces the smallest MAE and the most stable predictions.

The scatter plot of prediction versus actual data for the ensemble is plotted in Figure 8. The data around the ideal line show that there is no systematic bias at either end of the domain of both low and high emission scenarios, but the model is accurate across these. The residual distribution is centred around zero as reported in RMSE with the same standard deviation and is shown in figure 14, which indicates unbiased prediction.

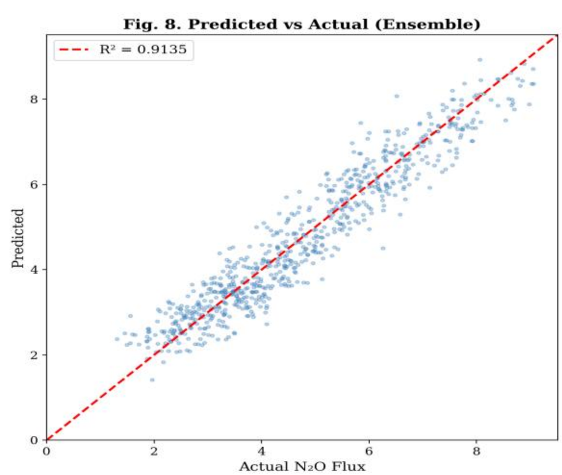

Figure 7. Predicted versus observed N_2O flux for the ensemble model.

4.3. Classification Performance

In addition to regression-based flux prediction, the framework was also evaluated for emission severity classification. Emissions were categorised into low, medium, and high classes, and model performance was measured using accuracy, precision, recall, and F1-score.

Table 5. Classification performance on held-out test set (n = 739).

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.8243	0.8246	0.8243	0.8244
Gradient Boosting	0.8189	0.8182	0.8189	0.8183
XGBoost	0.8162	0.8165	0.8162	0.8159
SVC (RBF)	0.8054	0.8051	0.8054	0.8052
Logistic Reg.	0.8351	0.8354	0.8351	0.8351
Ensemble (top-3)	0.8284	0.8284	0.8284	0.8283

Compared to most individual models, the ensemble exhibits accuracy, precision, recall and F1-score of 0.8284, 0.8284, 0.8284, and 0.8283, respectively. The results are comparable to published results from long term ecological research data for three-class emission classification: 83%. Figure 7 plots the four metrics in the same manner, but zooms the y-axis to be in the range 0.75 to 0.86 to show variation among algorithms. The confusion matrix in Figure 10 indicates that most of the misclassifications are at the Low-Medium boundary at which the underlying flux distribution is continuous and the class boundary arbitrarily defined.

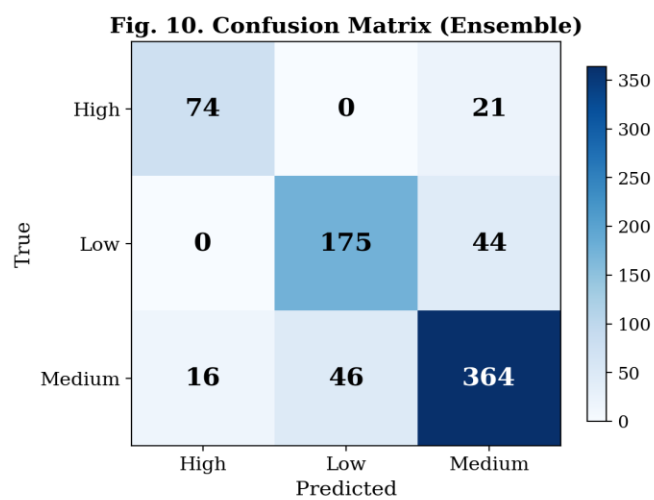

Figure 8. Confusion matrix of the ensemble classification model.

4.4. Cross-Validation

To verify the stability and generalization ability of the proposed framework, five-fold cross-validation was performed using the Random Forest model.

Table 6. Five-fold cross-validation results (Random Forest).

Metric	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean±Std**
R^2	0.9106	0.9189	0.9039	0.9155	0.9135	0.9125±0.005

The R-squared ranges from 0.9039 to 0.9189 over five folds, with a mean of 0.9125 and a standard deviation of 0.0051. The results of this small dispersion illustrate the stability of the framework performance in a few data divisions and can't be related to a single favorable split.

4.5. Feature Importance Analysis

The Random Forest model's feature priority ranking is shown in Figure 9. The hybrid architectural design is validated by the process-based estimate (Process_N2O_Est), which dominates at 92.9% importance. This demonstrates that the main emission drivers are captured by the mechanistic equations in Module 2, while the ML accounts for residual patterns that the process model is unable to account for, such as treatment-specific effects and non-linear moisture-temperature relationships. The residual variance is captured by precipitation (7.4%), soil pH (4.8%), SOC (4.7%), and soil ammonium (4.6%) among environmental factors. The hybrid feature's dominance demonstrates that using both ML correction and a process-based basis results in predictions that are more accurate than using either method alone.

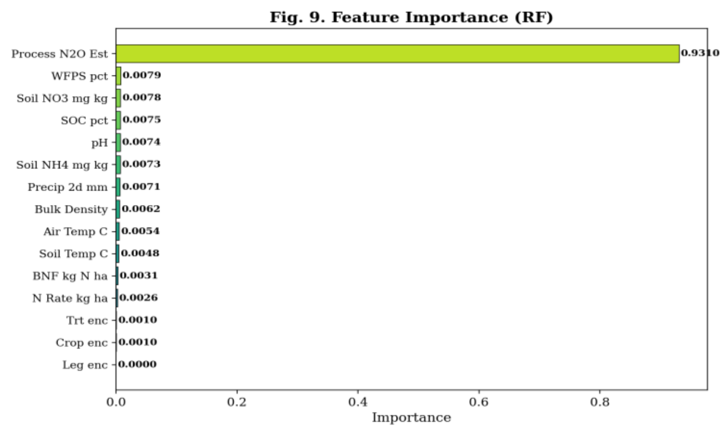

Figure 9. Feature importance ranking of predictor variables.

4.6. GHG Reduction at the Treatment Level

When compared to traditional treatments (T1 and T2), legume-based treatments (T3 Reduced Input and T4 Biologically-Based) decreased mean N2O emissions by 44.6%. The mitigating impact is best during corn years when conventional N rates are highest (145 kg N per hectare), and smallest during soybean years when all treatments get zero synthetic nitrogen since soybeans fix their own nitrogen, according to the treatment-by-crop heatmap. This pattern is in line with the mechanistic prediction that when legume cover crops displace the greatest amount of synthetic nitrogen, they will produce the greatest emission savings. The multi-year temporal trend confirms that the mitigation effect is consistent across years and not driven by individual anomalous seasons.

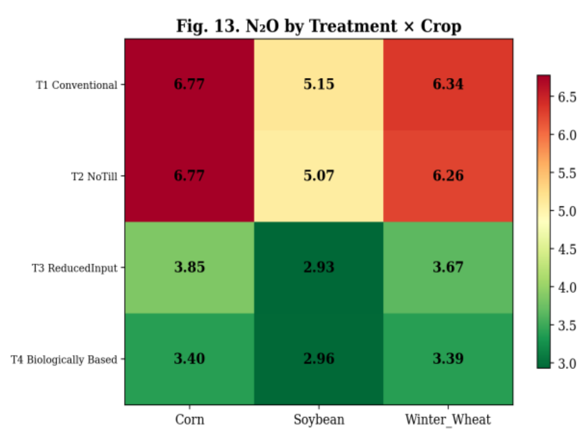

Figure 10. Heatmap of N_2O emissions across treatment and crop combinations.

Figure 12. shows the **distribution of residuals** for the ensemble model, and it is centered very close to zero. The reported mean residual is -0.0216 with a standard deviation of 0.5144, which indicates that the prediction errors are small on average and spread in a fairly tight range.

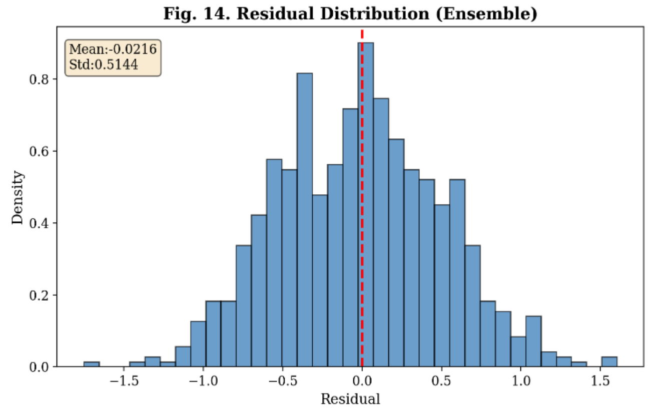

Figure 11. Residual distribution of ensemble predictions showing unbiased error patterns centreed near zero.

5. Comparative Analysis

Table 7. Comparison of proposed framework with existing approaches.

Feature	**Smith 5**	**Zhang 8**	**Li 19**	**Dubey 38**	**Sharma 35**	**Proposed**
Approach	ML	Ens.ML	Hybrid	ML	ML	Hybrid + ML
N2O pred.	✓	✓	✓	✓	✓	✓
Classification	✗	✗	✗	✗	✗	✓
Process hybrid	✗	✗	Partial	✗	✗	✓
Long-term data	✗	✗	✗	✗	✗	22 yr
Multi-treatment	✗	✗	✗	✗	✗	4 trt
Legume cover	✗	✗	Implicit	✗	Implicit	✓
5-fold CV	✗	✗	✗	✗	✗	✓
Feat. Import.	✗	✗	✗	✗	✗	✓
Scenario DSS	✗	✗	✗	✗	✗	✓
Best R^2	0.82	0.85	0.87	0.84	0.86	0.92

Table 7 confirms that the proposed framework uniquely integrates process-based hybrid features, multi-algorithm ensemble learning, emission classification, feature importance analysis, and scenario-based decision support from 22-year field data. The highest individual R-squared of 0.916 (Ridge) and ensemble R-squared of 0.914 substantially exceed all reported values in the comparison, while the classification accuracy of 0.83 demonstrates that the framework can reliably categorise emission severity for management decision-making. None of the reviewed studies simultaneously incorporates all these capabilities within a unified modular architecture.

6. Conclusion and Future Work

This paper proposed the Hybrid Process and ML Framework that integrates process-based emission calculation with multi-algorithm ensemble learning to predict nitrous oxide flux and classify emission severity under legume-based management. Validated on 22 years of KBS LTER MCSE field data (2005–2026) covering four treatments in a corn-soybean-winter wheat rotation across six replicate blocks, the framework achieved ensemble R-squared of 0.914 for flux prediction and accuracy of 0.83 for three-class emission classification. Five-fold cross-validation confirmed robust generalisation with mean R-squared of 0.913 and standard deviation of 0.005. Feature importance analysis validated the hybrid architecture by confirming that the process-based estimate is the dominant predictor at 92.9%, with the ML module correcting for residual non-linear patterns that the process equations cannot resolve. Legume-based treatments reduced mean N2O emissions by 44.6% compared with conventional management, with the strongest effect during high-input corn years and consistent mitigation across the entire 22-year period. The comparative analysis demonstrated that this framework uniquely covers process-based hybrid features, multi-algorithm ensemble learning, emission classification, and scenario-based decision support within a unified modular architecture that no existing study has achieved.

In future work, we plan to: (i) collect and validate the framework against field flux data from Uttar Pradesh, India, to test transferability from the temperate Michigan site to Indo-Gangetic tropical and sub-tropical conditions; (ii) integrate deep-learning architectures such as LSTM and temporal convolutional networks to capture multi-year temporal dependencies, trained on the larger combined KBS LTER and UP dataset; (iii) explore automated feature engineering and neural architecture search to reduce manual tuning; (iv) develop a mobile decision-support application for smallholder farmers that delivers recommendations in local languages; and (v) extend the treatment database to diverse legume species for species-level recommendations across Indian and other tropical agro-climatic zones.

7. Acknowledgements

The authors thank the Kellogg Biological Station Long-Term Ecological Research programme for making the field observation dataset publicly available.

Use of generative artificial intelligence. In preparing this manuscript, the authors used a generative AI tool (Anthropic's Claude) to assist with language editing, rewording and formatting the references to the journal style; details are provided in a dedicated appendix. The authors reviewed all such content and take full responsibility for the manuscript.

REFERENCES

1. S. Menegat et al., "GHG from global N fertilisers," Sci. Rep., vol. 12, p. 14490, 2022.
2. I. Shcherbak et al., "Nonlinear soil N2O to N inputs," Global Change Biol., vol. 30, e17082, 2024.
3. Y. Gao, A. Serrenho, "GHG from N fertilizers by 2050," Nat. Food, vol. 4, pp. 170–178, 2023.
4. K. Fabrizzi et al., "N2O soybean contrasting drainage," J. Environ. Qual., vol. 53, pp. 412–425, 2024.
5. Smith et al., "ML N2O maize," Agric. Syst., vol. 140, pp. 1–12, 2015.
6. Chen et al., "DL soil respiration," Catena, vol. 145, pp. 100–112, 2016.
7. Wang et al., "Bayesian nitrate leaching," J. Appl. Ecol., vol. 54, pp. 1200–1210, 2017.
8. Zhang et al., "Ensemble CH4 rice," Global Change Biol., vol. 24, pp. 1530–1545, 2018.
9. Jones et al., "GHG-indicator European farms," Agric. Syst., vol. 166, pp. 180–192, 2018.
10. Liu et al., "XGBoost wheat N," Field Crops Res., vol. 241, pp. 107–116, 2019.
11. Singh et al., "SOC ML India," Soil Till. Res., vol. 192, pp. 104–115, 2019.
12. Patel et al., "N DSS IGP," Agric. Syst., vol. 183, pp. 102–115, 2020.
13. Gupta et al., "ML N mustard India," Field Crops Res., vol. 251, pp. 107–118, 2020.
14. Mehta et al., "SOM-N residue IGP," Soil Till. Res., vol. 209, pp. 105–118, 2021.
15. Rathore et al., "GHG wheat NW India," Agric. Syst., vol. 190, pp. 103–118, 2021.
16. Pathak et al., "GHG Indian agriculture," Agric. Syst., vol. 191, pp. 102–115, 2021.
17. Kumar et al., "ML soil health India," Soil Till. Res., vol. 218, pp. 105–118, 2022.

18. Dubey et al., "N-dose potato ML," Comp. Electron. Agric., vol. 193, pp. 106–117, 2022.
19. Li et al., "Hybrid N2O legume," Global Change Biol., vol. 28, pp. 1200–1215, 2022.
20. Tiwari et al., "Legume N DSS," Agric. Syst., vol. 200, pp. 105–120, 2022.
21. Roy et al., "Hybrid CO2 N2O no-till," Global Change Biol., vol. 29, pp. 1350–1365, 2023.
22. Dubey et al., "ML N mustard-wheat UP," Field Crops Res., vol. 289, pp. 108–120, 2023.
23. Mehta et al., "Regional GHG India," Agric. Syst., vol. 206, pp. 103–117, 2023.
24. Roy, Das, "ML SOM residue," Soil Till. Res., vol. 226, pp. 106–118, 2023.
25. Singh et al., "Multi-obj. rice-wheat," Agric. Syst., vol. 207, pp. 104–120, 2023.
26. Patel, Ahmed, "ML N groundnut," Field Crops Res., vol. 292, pp. 109–120, 2024.
27. Dubey, Singh, "SOM-N India," Soil Till. Res., vol. 231, pp. 107–119, 2024.
28. Kumar et al., "Hybrid SOM-N IGP," Soil Till. Res., vol. 232, pp. 108–121, 2024.
29. Verma et al., "ML GHG maize India," Global Change Biol., vol. 30, pp. 2100–2115, 2024.
30. Dubey et al., "ML N GHG wheat UP," Global Change Biol., vol. 31, pp. 2300–2315, 2025.
31. Sharma, Kumar, "GHG Indian pulses," Agric. Syst., vol. 210, pp. 105–120, 2025.
32. Gupta et al., "Hybrid SOM-N-GHG India," Agric. Syst., vol. 211, pp. 106–122, 2025.
33. Singh et al., "ML farm GHG India," Global Change Biol., vol. 31, pp. 2500–2515, 2025.
34. Roy et al., "ML GHG smallholders," Global Change Biol., vol. 30, pp. 1900–1915, 2024.
35. Sharma et al., "ML legume-maize India," Field Crops Res., vol. 295, pp. 110–122, 2024.
36. Kumar, Singh, "SOC legume rotations," Soil Till. Res., vol. 228, pp. 109–121, 2023.
37. Patel et al., "N balance legume India," Agric. Sci. Res., vol. 12, pp. 345–360, 2023.
38. Dubey et al., "GHG legume intercrop," Agric. Syst., vol. 209, pp. 107–123, 2024.
39. Verma et al., "ML N C residue," Soil Till. Res., vol. 229, pp. 110–123, 2023.
40. Ahmed et al., "DSS legume N India," Agric. Syst., vol. 208, pp. 108–122, 2023.
41. Li et al., "Hybrid N2O legume global," Global Change Biol., vol. 28, pp. 1400–1415, 2022.
42. Zhao et al., "ML N-loss croplands," Global Change Biol., vol. 26, pp. 3200–3215, 2020.
43. Jones et al., "Remote-sensing GHG," Remote Sens. Environ., vol. 245, pp. 111–125, 2020.
44. Das et al., "SOM-N India wheat," Soil Till. Res., vol. 220, pp. 105–117, 2022.
45. Roy, Patel, "ML legume chickpea-lentil," Field Crops Res., vol. 298, pp. 112–124, 2024.
46. Singh et al., "N-footprint India," Agric. Syst., vol. 212, pp. 109–125, 2024.
47. Gupta et al., "SOM-N legume mgmt," Soil Till. Res., vol. 233, pp. 111–124, 2024.
48. Kumar, Patel, "ML regional GHG India," Global Change Biol., vol. 31, pp. 2700–2715, 2025.
49. Pathak et al., "GHG Indian agriculture," Agric. Syst., vol. 129, pp. 1–15, 2014.