Modelling N_2O Emissions in Fertilized Grasslands with Machine Learning: A Multi-Site LSTM Approach

Samin Payrosangari[1], Kathrin Fuchs[1], Benjamin Wolf[1], Ralf Kiese[1], Clemens Scheer[1]

[1]Karlsruhe Institute of Technology, IMK-IFU, Garmisch-Partenkirchen, Germany

Correspondence to: Samin Payrosangari, Kathrin Fuchs (saminpayro@gmail.com, kathrin.fuchs@kit.edu)

Abstract. Accurate quantification of GHG emissions from agricultural soils based on key environmental and management factors, is crucial for developing effective mitigation strategies. This study applies a machine learning approach using data from multiple montane grasslands in central Europe to model the spatial and temporal dynamics of N_2O emissions, using meteorological, soil, and management data. The primary aim is to advance predictive modelling of N_2O emissions from grassland soils for spatial upscaling and scenario analysis. Specifically, we assess whether a regionally trained model can accurately estimate N_2O emissions from an independent site excluded from training.

We collected data from five fertilized grasslands across southern Germany, northern Switzerland and western Austria (122 soil-site-treatment-year combinations), and trained a long short-term memory (LSTM) algorithm to model the influence of drivers within 5 subsequent days on N_2O emissions. The dataset includes daily N_2O emission measurements along with key emission drivers such as soil moisture and temperature in 10 cm soil depth, daily precipitation, occurrence of fertilization events (zero or one flag), as well as static soil characteristics such as pH and bulk density.

The trained LSTM model showed varying predictive performance when evaluated on two test datasets, with higher performance at some site-years and lower at others. One test set included temporally independent data (hold-out years) (RMSE = 28 µg N_2O-N m^{-2} h^{-1}), while the other included cross-site data from a soil type withheld from training (RMSE = 11 µg N_2O-N m^{-2} h^{-1}). The model captured N_2O dynamics, including the magnitude and timing of emission peaks driven by slurry application and environmental factors, though predictive performance varied across individual site-years. These results demonstrate that LSTM-based models can predict N_2O emissions for most site-years across test sites with similar environmental and soil characteristics, representing an alternative to process-based models for predicting soil N_2O emissions.

1 Introduction

Greenhouse gas (GHG) emissions significantly contribute to global warming, with agricultural activities accounting for approximately 12% of global anthropogenic GHG emissions (Chataut et al., 2023). The greenhouse gases emitted from soils are mainly methane (CH_4), nitrous oxide (N_2O) and carbon dioxide (CO_2), which are increasing due to high-energy-intensive agricultural production systems as well as use of mineral and organic fertilizers (Chataut et al., 2023). Today, agricultural activities are responsible for approx. two-thirds of total anthropogenic N_2O emissions and its high global warming potential makes it the third-most important anthropogenic greenhouse gas for global warming (Smith et al., 2007; Stocker et al., 2013). Thus, agricultural mitigation strategies play an important role in reducing N_2O emissions, which requires understanding the effects of management adoption (Fuchs et al., 2018).

The production of N_2O in the soil is driven by microbial processes — predominately denitrification and nitrification — which are influenced by both environmental conditions and management practices (Wang et al., 2021; Butterbach-Bahl et al., 2013). Key soil and environmental conditions regulating these processes include soil moisture, temperature, texture, nitrogen content, pH, and organic matter content, while fertilization, tillage, and irrigation represent the most important management practices (Wang et al., 2021; Butterbach-Bahl et al., 2013). The correlation between daily N_2O emissions and individual input drivers might be low and it's the interaction of drivers such as precipitation and temperature, as well as management practices that result in highly dynamic, variable and site-specific emission patterns (Wang et al., 2021). Thus, the complex influences of different drivers on emissions need to be simulated via high-performance models. This is especially relevant as direct measurement over large areas and long time periods is impractical (Butterbach-Bahl et al., 2013; Giltrap et al., 2010).

Modelling approaches for estimating N_2O emissions from agricultural systems include emission factors, statistical models, and process-based models (Butterbach-Bahl et al., 2013). Emission factors, i.e., a fixed proportion of applied nitrogen that is emitted as N_2O, are widely used for national inventory

reporting. However, they are unable to reflect the variability of soil and climate conditions that are relevant for N_2O emissions at the field scale or with a time resolution higher than yearly (Giltrap et al., 2010). In contrast, process-based models such as LandscapeDNDC (Haas et al., 2013) or Daycent (Parton et al., 1998) simulate the interactions between soil processes, climatic drivers and management. To this end, these models use a large set of parameters, which need to be parameterized based on calibration and validation datasets. As a consequence, uncertainty of model results due to parameterization is inevitable and the performance of such models heavily depends on the number and quality of calibration and validation datasets.

Given these limitations, machine learning (ML) offers a promising alternative or complementary approach to process-based modelling. ML models are trained on previous measurements and do not require a deep understanding or a definition of the underlying processes and parameters. However, implementing a high-accuracy ML model comes with challenges, such as the need for large, high-quality datasets and expert knowledge about the used algorithms, along with lower interpretability than process-based models. While implementing the right ML model requires effort, once a generalized model is built, applying it to new data is highly time-efficient.

Building an ML model for estimating N_2O emissions requires considering short-term fluctuations and peaks in N_2O emissions in response to temporal dynamics in soil moisture and temperature, such as dry–wet transitions, water logging, as well as management activities (Butterbach-Bahl et al., 2013). Therefore, estimating daily values of N_2O emissions can be improved by considering the changes in drivers over several days prior to the emission event, which is addressed in this study. Long short-term memory network or LSTM (Hochreiter and Schmidhuber, 1997) is a machine learning model that detects the dynamic relationships in sequential data. It has previously been used for different applications such as flood forecasting (Le et al., 2019), predicting water table depth in agricultural areas (Zhang et al., 2018), as well as modelling soil N_2O and CO_2 emissions (Hamrani et al., 2020).

In this study, we developed an LSTM model to simulate the daily N_2O emissions of grasslands in central Europe, with the three main objectives:

1) to train an LSTM model using data from multiple soils and climates to simulate daily N_2O emissions at alpine grasslands
2) to test the model performance on an independent grassland soil not used for model training to assess the regional applicability of the model to pre-alpine grasslands without site-specific parametrization
3) to assess the relative importance of different environmental and management variables for predicting N_2O emissions

2 Materials and methods

Daily and growing season N_2O emission data were collected from multiyear field experiments of 5 grassland sites representing a variety of soils at different elevations and thus climate conditions in Central Europe: South Germany (Fendt, Rottenbuch and Graswang), west Austria (Neustift) and North Switzerland (Chamau). Table 1 shows the characteristics of these grasslands, all of which are managed with organic fertilizer (manure and slurry) and cutting. While N_2O emissions were highest at Chamau, lowest N_2O emissions were observed at the Neustift site. The collected dataset for training the LSTM model includes daily measurements of soil N_2O emissions, soil physical and chemical properties (Table 2), as well as daily meteorology. Daily N_2O fluxes were always expressed as N_2O-N in this paper.

Grassland	Elevation (m a.s.l.)	Annual N_2O-N (kg ha^{-1} y^{-1})	Annual precipitation (mm)	Avg annual air temperature (°C)	pH	BD (g cm^{-3})
Chamau Switzerland (CH-Cha)	393	5.28	1179	9.8	6.5	0.94
Fendt Germany (DE-Fen)	595	2.83	962	9.3	4.8-6.5	0.8-1.1
Rottenbuch Germany (DE-RbW)	769	1.69	1047	8.6	4.8-5.7	0.7-0.95
Graswang Germany (DE-Gwg)	864	2.55	1464	7	5.1-6-6	0.52-0.78
Neustift Austria (AT-Neu)	970	0.03	852	6.5	5.7	1.2

Table 1 Summary of the properties of grasslands in different climates and elevation used in this study (Kiese et al., 2018; Soltani et al., 2018; Gilgen & Buchmann, 2009; Roth, 2006; Hörtnagl et al., 2014). The total annual N_2O-N is averaged over different years of data availability for each grassland.

2.1 Site descriptions

German Terrestrial Environmental Observatories (TERENO) Pre-Alpine

The Tereno Pre-Alpine observatory includes three grassland sites, Graswang (DE-Gwg), Rottenbuch (DE-RbW), and Fendt (DE-Fen), at elevations of 864, 769, and 595 m a.s.l. respectively, which are located in the Bavarian Prealps in southern Germany (Kiese et al., 2018; Soltani et al., 2018). Automatic chamber measurements of N_2O emission along with a wide range of other measurements such as soil moisture and temperature are conducted by the Karlsruhe Institute of Technology (KIT). The automatic chamber measurements are carried out on 36 lysimeters, i.e., soil cores with 1.0 m^2 surface and a depth of 1.4 m (Kiese et al., 2018). A subset of 18 out of 36 soil cores are translocated from the high-elevation site Graswang and the medium-elevation site Rottenbuch to the lower altitude sites (space-for-time-approach), referred to as Graswang in Fendt, Graswang in Rottenbuch or Rottenbuch in Fendt translocations to investigate the soils under climate change scenarios (Kiese et al., 2018). The lysimeter setup for these grasslands is shown in Fig. A1 of appendix A. Experiments on each lysimeter follow intensive or extensive management treatments, where extensive management includes up to three cuts and two manure applications, and intensive management includes up to six cuts and five slurry applications per year (Kiese et al., 2018). From each soil and treatment three replicates are considered, for instance, there are three Graswang in Fendt soil core replicates which are intensively managed as shown in Fig. A1. Notably, at the time of this study, sub-daily (up to 6 flux rates per day) data from 24 lysimeters operated in Fendt (Fig. A1 a) and Graswang (Fig. A1 c), which are aggregated (averaged) into daily are used according to availability of data for different years between 2014 and 2020. The soil properties of different replicates (N=3: L1-L3) are shown in Table 2.

Switzerland FLUXNET Chamau

Chamau (CH-CHA) is a permanent grassland site that was established by ETH Zurich in 2005 (Feigenwinter et al., 2023) and is part of Swiss FluxNet initiative. This site is located in the Reuss River valley on the Swiss Plateau, approximately 30 km south of the city of Zurich, Switzerland, at an elevation of 393 m a.s.l. Half-hourly eddy-covariance measurements of N_2O and CH_4 started at this site in 2012 (Merbold et al., 2014). Moreover, a N_2O mitigation experiment, from 2015 to 2020, was conducted by establishing separate parcels in the northern and southern parts of the site, continuing slurry fertilization on the control parcel in the south and oversowing the north parcel with a clover seed mixture without fertilization (Fuchs et al., 2018). The non-gap-filled measurements from the control parcel aggregated (averaged) into daily values are used in this study, which represent an intensively managed site with cutting 4–5 times a year followed by liquid slurry application plus one slurry application in spring each year (Feigenwinter et al., 2023; Hörtnagl et al., 2025; Fuchs et al., 2018).

Austria Neustift

The Neustift grassland (AT-NEU) is an intensively managed meadow in the middle of a flat valley at the bottom of the Stubai valley in the Austrian alps, at an elevation of 970 m a.s.l (Wohlfahrt et al., 2008; Wohlfahrt et al., 2012; Hörtnagl & Wohlfahrt, 2014). The N_2O and CH_4 emissions at this site were measured between April 2010 and February 2012, and the management consisted of three cuts and one solid manure and slurry application in autumn each year, which was considered an intensively managed site (Wohlfahrt et al., 2012; Hörtnagl & Wohlfahrt, 2014; Hörtnagl et al., 2018). At this site, fluxes were calculated via virtual disjunct eddy-covariance, providing half-hourly gap-filled flux measurements (Wohlfahrt et al., 2008; Wohlfahrt et al., 2012; Hörtnagl & Wohlfahrt, 2014), which we averaged to daily values for this study.

Climate (Grassland)	Replicate-Management	Clay 0-10 cm (%)	Silt 0-10 cm(%)	Sand 0-10 cm (%)	SOC 0_10 cm (%)	Nt 0_10 cm (%)	C/N 0_10 cm	pH 0_10 cm	Bulk density 0_10 cm (gr cm^{-3})	Fertilizer (Kg N ha-1 y-1)	Dataset assignment	Data availability years	Data size
Graswang	L1-Intensive	55	42	3	11.4	1.2	9.48	6.40	0.52	187.00	Train	2020	277
	L2-Intensive	38	43	19	8.7	0.81	10.74	5.43	0.73	187.00	Train	2020	273
	L3-Intensive	59	35	6	7.9	0.83	9.53	6.02	0.68	187.00	Train	2020	269
	L1-Extensive	60	35	5	8.5	0.87	9.72	6.07	0.65	70.00	Train	2020	284
	L2-Extensive	58	39 0	3	11.1	1.13	9.86	6.45	0.55	70.00	Train	2020	284
	L3-Extensive	37	44	19	8.1	0.75	10.80	5.20	0.80	70.00	Train	2020	283
Fendt	L1-Intensive	61	36	3	10.9	1.15	9.43	6.56	0.55	187.00	Train/Temporal test	2014-2019/2020	744/172
	L2-Intensive	37	44	19	7.5	0.73	10.27	5.35	0.78	187.00	Train/Temporal test	2014-2019/2020	979/224
	L3-Intensive	61	35	4	8.5	0.9	9.44	6.12	0.66	187.00	Train/Temporal test	2014-2019/2020	1005/258
	L1-Extensive	57	40	3	11.3	1.195	9.42	6.40	0.52	76.00	Train/Temporal test	2014-2019/2020	1076/245
	L2-Extensive	58	38	5	7.7	0.81	9.51	6.12	0.53	76.00	Train/Temporal test	2014-2019/2020	915/238
	L3-Extensive	38	43	20	8.0	0.82	9.76	5.10	0.77	76.00	Train/Temporal test	2014-2019/2020	1039/213
	L1-Intensive	25	40	35	5.2	0.52	9.92	5.87	0.93	187.00	Train/Temporal test	2014-2019/2020	915/228
	L2-Intensive	30	43	27	4.5	0.45	9.89	4.87	1.01	187.00	Train/Temporal test	2014-2019/2020	876/260
	L3-Intensive	33	46	21	6.4	0.68	9.35	5.16	0.78	187.00	Train/Temporal test	2014-2019/2020	1095/252
	L1-Extensive(2014-2018), Intensive(2019-2020)	27	42	30	5.5	0.57	9.72	5.72	0.86	97.00	Train/Temporal test	2014-2019/2020	1012/240
	L2-Extensive(2014-2018), Intensive(2019-2020)	31	42	28	4.4	0.45	9.69	4.90	0.96	97.00	Train/Temporal test	2014-2019/2020	714/258
	L3-Extensive(2014-2018), Intensive(2019-2020)	34	45	21	6.4	0.69	9.32	5.26	0.81	97.00	Train/Temporal test	2014-2019/2020	751/255
Chamau	Intensive	19	45	36	3.2	0.33	9.79	6.50	0.94	204.00	Train/Temporal test	2013, 2014, 2016, 2019,2020/ 2015	1420/ 320
Neustift	Intensive	4	43	53	3.0	0.28	10.71	5.70	1.20	340.00	Train	2010-2011	259
Fendt	L1-Extensive(2014-2018), Intensive(2019-2020)	40	41	19	7.0	0.725	9.66	5.73	0.69	97.00	Cross-site Test	2014-2020	1320
	L2-Extensive(2014-2018), Intensive(2019-2020)	24	42	34	4.3	0.44	9.75	5.34	0.95	97.00	Cross-site Test	2014-2020	1452
	L3-Extensive(2014-2018), Intensive(2019-2020)	23	53	24	5.8	0.63	9.19	5.43	0.81	97.00	Cross-site Test	2014-2020	1478
	L1-Intensive	24	43	33	4.4	0.47	9.32	4.84	0.93	187.00	Cross-site Test	2014-2020	1554
	L2-Intensive	36	40	24	6.2	0.63	9.84	5.39	0.70	187.00	Cross-site Test	2014-2020	1532
	L3-Intensive	28	49	23	5.9	0.65	9.08	5.18	0.80	187.00	Cross-site Test	2014-2020	1354

Table 2 Soil properties of the different soils in various climates used in this study (Kiese et al., 2018; Soltani et al., 2018; Gilgen & Buchmann, 2009; Roth, 2006; Hörtnagl et al., 2014; Hörtnagl et al., 2018; Feigenwinter et al., 2023). All soil properties refer to measurements in 0-10 cm depth. Dataset size refers to the number of daily flux observations.

2.2 Input variables and model setup

In this study we model the average daily N_2O-N emissions based on the temporal and static drivers available in the collected dataset. The temporal drivers used by the model are the daily averages of soil moisture and temperature at 10 cm soil depth, total precipitation, and the occurrence of a fertilization event. The model also incorporates static inputs, such as soil pH and bulk density, to represent the soil's chemical and physical properties across the different sites.

To estimate the daily N_2O-N emission, the sequential LSTM model walks through the data sequences and in each round processes a subsequence of length 5 which include the information as Eq. 1 for estimating average daily fluxes:

$$\begin{Bmatrix} Soil\ mois_{t-4} & Soil\ temp_{t-4} & Precipitation_{t-4} & Fertilization_{t-4} \\ Soil\ mois_{t-3} & Soil\ temp_{t-3} & Precipitation_{t-3} & Fertilization_{t-3} \\ Soil\ mois_{t-2} & Soil\ temp_{t-2} & Precipitation_{t-2} & Fertilization_{t-2} \\ Soil\ mois_{t-1} & Soil\ temp_{t-1} & Precipitation_{t-1} & Fertilization_{t-1} \\ Soil\ mois_{t} & Soil\ temp_{t} & Precipitation_{t} & Fertilization_{t} \end{Bmatrix}, \{pH, BD\} \rightarrow \{N_2O - N_t\} \quad (1)$$

Temporal Static Target

The indices indicate the values of the features at different days in which day t is the current day. The average daily N_2O-N emissions at day t is the target of the machine learning model.

The implemented sequential LSTM models in this study rely solely on the temporal and static input features shown in Eq. 1 in both training and testing phases, and no information from the N_2O-N emissions on previous days is used for learning the N_2O-N emission of the current day.

2.3 Dataset structure and pre-processing

The availability of a diverse dataset that includes translocated soil cores across German grasslands, also using data from Switzerland and Austria gives this study the opportunity to cover a variety of soils in different climates (Table 2), which is key in creating a regional model capable of transferring results across sites.

At German sites, replicates sampled from individual grassland often exhibit varying soil properties, as they are taken from different sections of expansive grassland areas. Therefore, for training the model the data from each replicated lysimeter was used separately so that the model could learn the role of soil properties in more detail.

We split the dataset based on temporal hold-out as well as cross-site approaches to assess the transferability of the model in temporal and spatial scale. The training set includes treatments with a variety of soil properties, N_2O emission patterns and climates which includes (1) several years from Chamau, Fendt and Graswang soils with intensive and extensive treatments and (2) the full dataset from the Neustift site due to its limited sample size. Neustift represents the highest-altitude grassland with the lowest emissions, low average temperature and moisture and therefore provides important information for training the model. The training set includes around 14500 data rows in total. During the training process, the training dataset partitions into training and validation parts following a random 80:20 split as a requirement for machine learning model training. The training partition is used to fit the model, while the validation partition facilitates hyperparameter optimization—such as learning rate decay—and helps monitoring model's progress during training.

On the other hand, the test set which was completely unseen during the training, was chosen according to the study's objectives:1) Temporal test set: One year of data from Chamau, and all treatments of Fendt and Grasswang soils is held back from the training to be used as a test set. This demonstrates the model's ability to predict data from unseen years, which is consistent with the temporal training/testing split used for time series forecasting. This test set includes 3,160 data rows, (2) cross-site test set: This test set is for spatial evaluation and utilises all available Rottenbuch soil data. The cross-site test set encompasses approximately 8,690 data rows.

Along with properly splitting the dataset, data cleansing and pre-processing is crucial for training the ML model. In this study the quality checked measured data from which the gaps were excluded are used. In addition, scaling the data prior to training ensures faster convergence and improves model performance by preventing features with larger magnitudes from disproportionately influencing the learning process. The input drivers are all scaled via the Scikit-learn library robust scaler (Pedregosa et al., 2011) shown in Eq. 1. The choice of using robust scaler is made after investigating other options such as standardization and normalization (min–max scaler) for all or individual drivers according to their distribution. The robust scaler transforms each feature by removing the median and scaling it according to the interquartile range, making it robust to outliers. For each variable x, the scaled value x_{scaled} is computed via Eq. 2:

$$x_{scaled} = \frac{x - Median(x)}{Q_3(x) - Q_1(x)} \quad (2)$$

Moreover, the target value (daily N_2O-N emission) is log-transformed via natural logarithm (ln) for model setup, as its original values show a highly right-skewed distribution as shown in Fig. 1. Therefore, the output of the model is generated at log-scale and subsequently needs to be back transformed to the actual emission scale via the exponential function. Logarithmic transformations are widely used to reduce skewness and stabilize variance in heteroscedastic data (West, 2021; Hastie et al., 2009; Box & Cox, 1964). Large differences in gradient magnitude and high stochastic gradient variance are known to impair optimization and cause slow convergence in deep neural networks (Goodfellow et al., 2016); hence, variance-stabilizing transformations can yield better-conditioned gradients, more stable training and faster convergence (Salinas et al., 2020; Smyl, 2020).

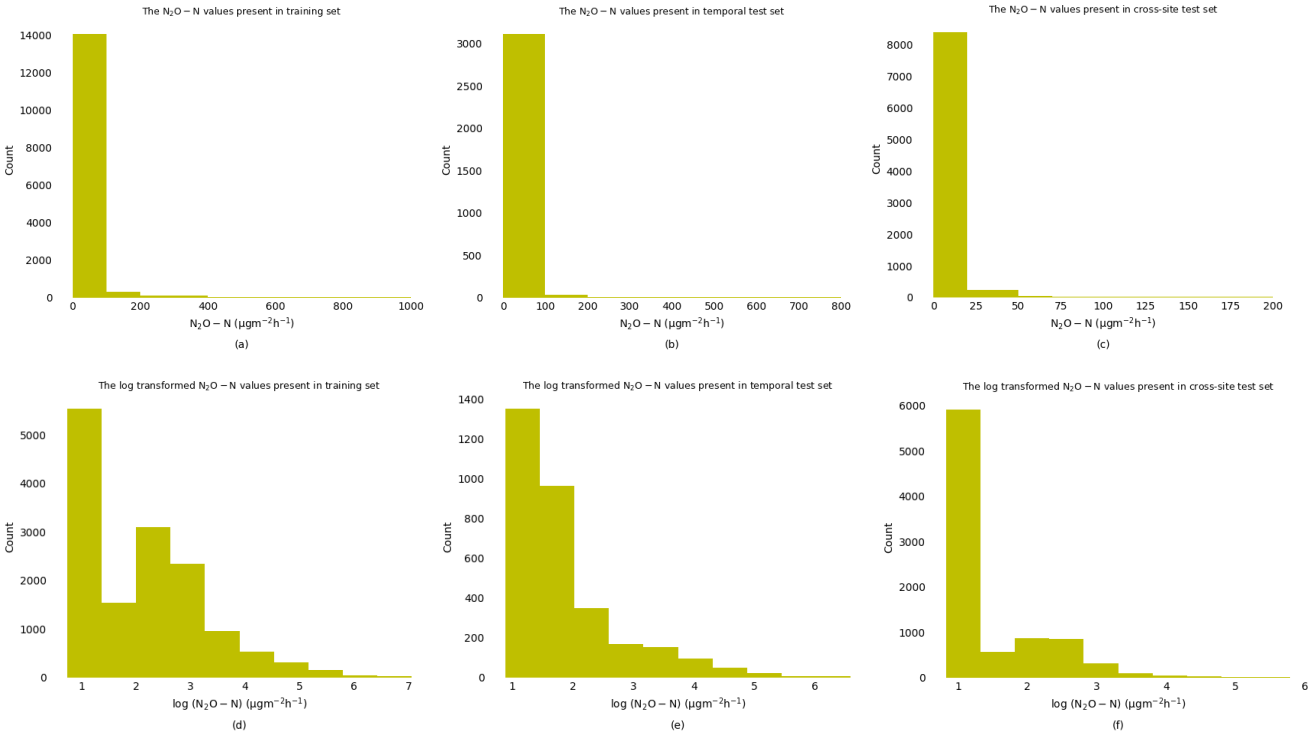


Figure 1 Histogram of original and log-transformed daily N_2O-N emissions values in the dataset used in this study. Panels a-c show the original values, and panels d-f show the log-transformed (natural log) values for the training set, temporal test set and cross-site test sets respectively.

2.4 Model development and training procedure

To build a model based on LSTM neural networks with high performance in simulating daily N_2O-N emissions, different aspects such as the architecture of the neural network, scaling and transforming the data, selecting a proper loss function for training and hyperparameter tuning were considered. In this section, the methods used in this study for model development are explained.

2.4.1 LSTM model architecture

Long short-term memory (LSTM) is a type of recurrent neural networks (RNN) that learns long-term dependencies in sequential data and is often used in speech recognition and time series processing (Hochreiter and Schmidhuber, 1997). As discussed in previous sections, simulation of daily N_2O-N emissions from soil is more accurate if the role and changes in environmental conditions over several days are simulated using a sequential model. In this study, we built the LSTM network architecture as shown in Fig. 2, which was implemented using the Keras library in Python (Chollet et al., 2015). The model architecture is investigated with various configurations and hyperparameters to find the optimum model.

The models in this study are built with an attention mechanism. Attention in LSTM refers to a mechanism that allows the model to dynamically focus on specific parts of the input sequence when generating each output, improving performance by weighting relevant hidden states (Bahdanau et al., 2015).

Figure 2 shows a brief representation of the best model's architecture. The complete architecture of the model is available at the Github link provided in Data and code availability section. The model initially processes the temporal inputs (Eq. 1) via 3 LSTM layers (Fig. 2). Subsequently, an attention layer is applied to the full sequence of hidden states produced by the last LSTM layer, allowing the model to learn weighted dependencies across all time steps. The resulting attention representation is added back to the original LSTM outputs through a residual connection, and the final time step of this enhanced

temporal representation is concatenated with the static features to proceed toward the output (Fig. 2). The mathematical formulation of the LSTM architecture and the attention mechanism are described in detail in the original publications (Hochreiter and Schmidhuber, 1997; Bahdanau et al., 2015).

In problems with both temporal and static features, there is an alternative possible architecture which initially processes the static inputs via dense layers, then uses the processed information as an initial input to LSTM layers (initializing the hidden state of first LSTM layer). However, since this alternative architecture showed low performance for our dataset, we did not further investigate it.

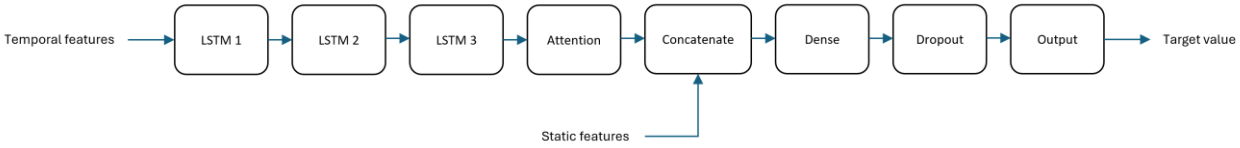


Figure 2 Architecture of the LSTM model used in this study. The LSTM implementations is available in the repository provided in the code availability section.

2.4.2 Loss function

The loss function used during training quantifies the error between the model's predictions and the actual target values, guiding the optimization process to minimize this error. For the current imbalanced dataset, a custom loss function is used to highlight the importance of catching high peaks of N_2O-N emissions which is formulated in Eqs. (3) and (4) as:

$$loss = abs(y_{true} - y_{pred}), \tag{3}$$

$$weighted\ loss\ = loss * e^{loss_coefficient * loss}, \tag{4}$$

With y_{true} and y_{pred} representing the observed and predicted daily N_2O-N fluxes, respectively. This loss function is significantly penalizing the model for high errors in peak values and showed a better performance comparing to the mean squared error (MSE), mean absolute error (MAE) as well as weighted mean squared (WMSE) error. In addition, loss_coefficient is a hyperparameter controlling the exponential weight further, with potential values shown in Table 3.

2.4.3 Hyperparameter tuning

Hyperparameter tuning is a key step in developing machine learning models, as it helps identifying the best-performing configuration for a given problem. Hyperparameters influence the model's complexity, learning ability, and performance. The main hyperparameters used in this study and their tested ranges are shown in Table 3. In this study, Hyperband algorithm (Li et al., 2018) was used at initial stages to optimize the hyperparameters of the LSTM model. Hyperband improves upon conventional random search by efficiently balancing exploration and exploitation, initially evaluating a broad range of hyperparameter configurations and progressively allocating more computational resources to the most promising trials (Li et al., 2018). In addition, manual adjustments were applied based on dataset characteristics to improve the model's ability in capturing fluctuations and rare high peaks more accurately.

Hyperparameter	Value ranges
Learning rate	[1e-2, 1e-3]
Activation functions for dense layers	[Tanh, Relu, Leaky relu, Linear]
Activation functions for LSTM layers	[Tanh, Relu]
Number of neurons	20-65
LSTM layers recurrent L2 regularization	[1e-4,1e-3,1e-2]
LSTM layers kernel L2 regularization	[1e-4,1e-3,1e-2]
Dense layers L2 regularization	[1e-4,1e-3,1e-2,1e-1]
Dropout	[0.01,0.05,0.1, 0.2,0.5]
Number of LSTM Layers	2-5
Number of dense layers	2-3
loss_coefficient used in the loss function	[1,1.2, 1.5]

Table 3 Ranges of hyperparameters of the model considered for hyperparameter tuning in this study. The best hyperparameters are available in the repository provided in Data and code availability section.

2.4.4 Feature importance in LSTM

This study applies the permutation feature importance (PFI) method introduced by Breiman (2001), which calculates the contribution of each feature to the model's estimation of N_2O-N emissions. This technique shuffles the value of an individual feature in the dataset, calls the model and observes the degradation of the model performance, which is now producing results with this feature having shuffled values (Breiman, 2001). Therefore, we define the relative importance metric of each feature as the degree of increase in the mean squared error (MSE) in comparison to the MSE in the original model, formulated in Eq. 5:

$$\textit{Relative importance feature } m = \textit{MSE PFI feature } m - \textit{MSE original model} \quad (5)$$

2.5 Post-processing and statistical analysis

We report the coefficient of determination R^2 (Montgomery et al., 2021), the mean absolute error MAE and root mean squared error RMSE (Willmott & Matsuura, 2005), as well as bias (mean difference between modelled and observed values) to indicate the model's performance. In addition, we report Spearman's rank correlation coefficient ρ and its 95% confidence interval (Spearman, 1904; Bonett and Wright, 2000) to capture monotonic relationship between modelled and observed values. Several metrics for interpreting the performance have to be used in combination, as some of the metrics such as R^2 are sensitive to outliers and few extreme errors could lead to misleading accuracy interpretation with them (Khoshvaght et al., 2025). In addition, for the analysis of the results time series plots for daily emissions as well as cumulative plots for yearly cumulative emissions are created in Matplotlib, Python (Hunter, 2007).

To report the results of different grassland systems in Germany, the model's daily N_2O-N estimations of individual soil replicates are generated and are finally aggregated on replicates of same treatment.

3 Results

3.1 Comparison of different model configurations

Table 4 compares the RMSE of some of the experiments conducted in this study to identify the optimal model configuration, in which different subsets of soil properties or length of input sequences (5, 10, 20) were tested. For each experiment the hyperparameter tuning for the LSTM model is done extensively to obtain an optimum model fitting to the input. In the current dataset, Bulk density (BD) showed strong negative Pearson correlation of -0.9 to both soil organic carbon (SOC) and clay. While all these variables are key drivers of N_2O emissions, only bulk density was included in the model due to slightly better performance in capturing emission peaks. Adding correlated features reduces model's accuracy, likely due to confusing the model with redundant features which do not provide extra information. The results presented in next section correspond to the best-performing model in catching low and high emissions as well as cumulative emissions (index 1 in Table 4).

Index	Input soil properties	Test RMSE ($\mu g m^{-2} h^{-1}$)	Sequence length
1	**BD, pH**	**18**	**5**
2	BD, pH	23	10
3	pH, SOC	19	5
4	BD, pH	19	20
5	SOC, BD, Clay, pH	20	5

Table 4 Model specifications of 5 experiments compared in this study. The models are compared on the basis of the RMSE on the temporal test set and cross-site test set (averaged), the static inputs they receive, and the length of the input sequence. The temporal drivers are identical between the experiments.

3.2 Model performance

The best model performed on the training set with RMSE of 34 $\mu g\ m^{-2} h^{-1}$, R^2 of 0.51 and a low negative bias of -5 $\mu g m^{-2} h^{-1}$ (Table 5). Figure 3 Shows the scatter plots of model's performance on the training, temporal and cross-site sets. In this section we discuss the performance of the model in details.

	Log scale						Original scale					
	R2	**Spearman ρ**	**Spearman_CI 95%**	**RMSE**	**MAE**	**Bias**	**R2**	**Spearman ρ**	**Spearman_CI 95%**	**RMSE**	**MAE**	**Bias**
Train set	0.60	0.65	[0.64, 0.66]	0.65	0.54	0.013	0.51	0.65	[0.64, 0.66]	34	10	-5
Temporal test	0.35	0.57	[0.55, 0.58]	0.70	0.54	0.32	0.30	0.57	[0.55, 0.58]	28	9	3
Cross-site test	-0.09	0.38	[0.36, 0.39]	0.70	0.58	0.36	-0.12	0.38	[0.36, 0.39]	11	5	2

Table 5. Statistics for train, temporal test and cross-site test datasets on the log scale and the original scale (index 1 in Table 4). RMSE, MAE and bias values in original scale are reported in $\mu g\ m^{-2} h^{-1}$ unit.

As discussed in previous sections, model training and evaluation were conducted in log space to stabilize variance and prevent extreme observations from dominating error metrics. While log-transformation improves the model's performance, it makes the interpretation of the results more challenging (Hodson, 2022). Consequently, performance metrics are reported on both the log scale and the original scale (Table 5). In the following sections, we primarily discuss results on the original scale to better reflect predictive behaviour.

Temporal test set model performance

To evaluate temporal generalization within sites, model performance was assessed on hold-out years not used for training. For the temporal test set the Spearman's rank correlation (ρ), reflecting the strength of the monotonic relationship between observed and predicted values, was 0.57 (95% CI [0.55–0.58]) indicating robust generalization with no evidence of overfitting (Prion and Haerling, 2014). The narrow confidence intervals across datasets among all training, temporal test as well as cross-site test sets demonstrate that these estimates are statistically stable and precise (Ruscio, 2008). Despite some performance degradation comparing to the training set, the preserved rank correlation shows that the model reliably captures the temporal ordering of emission events across hold-out years in many cases, while it shows a lower performance at others such as Fig 4 c.

The R^2 value representing the proportion of explained variance, was 0.30 for the temporal test set, lower than 0.51 of the training data (Table 5). This indicates that a fraction of observed variability is retained under temporal extrapolation. The reduction in R^2 indicates a moderate loss of predictive skill on test data, which is expected given the highly episodic nature of N_2O emissions and the intentionally limited set of input variables used in this study.

Bias for the temporal test set was +3 μg m^{-2} h^{-1}, indicating a slight overall overestimation. This systematic deviation is small and suggests minor systematic errors under temporal extrapolation. The RMSE of 28 μg m^{-2} h^{-1} quantifies the average prediction error, with greater weight assigned to large deviations. Together with an MAE of 9 μg m^{-2} h^{-1}, these metrics indicate moderate errors.

The model captures the temporal variability of N_2O emissions and, in many cases their magnitude, although performance varies across sites (Fig. 4). At Fendt in Fendt 2020 intensive, the model predicted the overall variability of the emissions very well (Fig. 4 b). At Chamau, the model captures the fact that this soil experiences higher peaks than other soils, while individual peak magnitudes were occasionally over- or underestimated or slightly delayed leading to lower performance metrics (Fig. 4 a). At Graswang in Fendt soils, fluxes were more frequently overestimated (Fig. 4 c, d), showing a significantly lower performance and higher bias comparing to Fig 4. b. Deviations in peak magnitude and timing account for the reduction in R^2.

Overall, the temporal test results demonstrate generalization across hold-out years in several cases. While some reduction in explained variance and peak accuracy occurs, the preserved rank correlation, narrow confidence interval, positive R^2, and low bias indicate levels of temporal transferability, which could be improved with extending the dataset further with more data.

Cross-site model performance

The cross-site evaluation is based on an independent dataset, comprising a soil (Rottenbuch in Fendt) not used during model training, thereby providing a reliable test of spatial generalizability. Spearman's rank correlation (ρ) for the cross-site set was 0.38 (95% CI [0.36–0.39]), indicating a moderate, statistically precise monotonic relationship which is lower than the observed value for the temporal test set (Table 5). Although reduced, the correlation remains statistically robust indicated by the narrow confidence intervals. The mean bias across the cross-site test set was +2 μg m^{-2} h^{-1}, indicating a slight overall tendency toward overestimation and a magnitude comparable to that observed for the temporal test set. The overall RMSE of 11 μg m^{-2} h^{-1} and MAE of 5 μg m^{-2} h^{-1} reflect the average magnitude of prediction errors which is also strongly influenced by a few extreme misestimations.

Figure 4 e-n show the model's performance on the cross-site test set. For a more detailed assessment of performance on the cross-site test set, performance metrics were calculated separately for each site-year and are reported in the titles of the individual panels in Fig. 4. The R^2 values ranged from 0.24 to 0.64 for six site-years (Fig. 4 e, f, g, h, j, and k), indicating a moderate to good agreement between observed and predicted emissions. In contrast, four site-years (Fig. 4 i, l, m, and n) exhibited substantially lower R^2 values, ranging from −25 to 0.14. These results demonstrate that the overall R^2 value of −0.12 reported for the cross-site test set in Table 5 does not reflect a uniformly poor model performance or a systematic failure within the whole cross-site test set. Rather, it is strongly influenced by a small number of challenging site-years that disproportionately affect the aggregated metric.

A proportion of the reduced R^2 at the challenging site-years is caused by the few extreme prediction errors at fertilization events, while the rest is caused by the tendency toward overestimation during these years shown by the high bias values (Fig. 4 i, l, n).

Substantial overestimations were observed at Rottenbuch in Fendt 2017 and 2020 intensive management campaigns following fertilization (Fig. 4 i, l, n). In 2015, 2016, and 2019 fertilization events occurring during warm periods were associated with pronounced emission peaks even exceeding 150 µg m^{-2} h^{-1} (Fig. 4 h, j, k). Despite similar temperature and moisture, the 2017 and 2020 campaigns did exceptionally not exhibit elevated N_2O fluxes after fertilization. The model nevertheless predicted several large emission peaks after fertilization. This results in the extreme prediction errors at these site-years contributing to the negative R^2 (Fig. 4 i, l, n). For instance, excluding the three extreme outliers from Rottenbuch in Fendt 2017 intensive in Fig. 4 l, resulted in R^2 increasing from −15.36 to −0.51. However, the bias value remains high, equal to 4.10, indicating that the model shows levels of overestimations also for other days at this challenging site-year, contributing further to the negative R^2 (Fig. 4 l).

Even for site-years with low R^2 values, the relatively high Spearman correlations, 0.42-0.61, indicate that the model was generally able to reproduce parts of the temporal dynamics of N_2O emissions (Fig. 4 i, l, n).

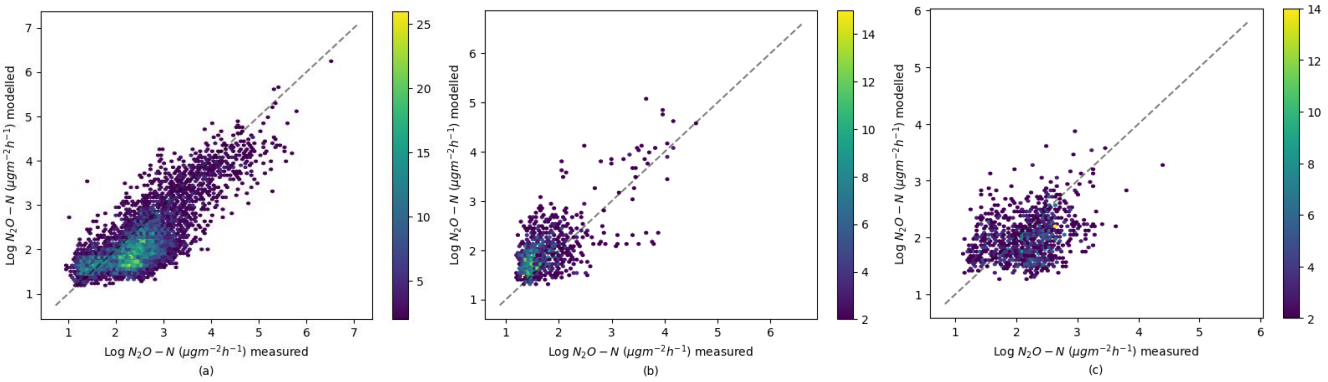


Figure 3 Hexbin plots with density colour bar comparing observed (x-axis) with simulated (y-axis) daily N_2O-N fluxes on a natural log (ln) scale for (a) the training set and (b) the temporal test set, (c) cross-site test soils of Rottenbuch in Fendt. Hexagon colors show the number of predictions falling into them, appearing as lighter (yellow) colors for bins with more data points and darker colors (blue) with less data points.

Cumulative emissions

In addition to the analysis of model performance for daily emissions, the cumulative emissions were assessed. Figure 5 shows the comparison between observed and modelled annual cumulative emissions achieved by aggregating the available daily fluxes. The results show that the model creates an accurate estimation of annual emissions for some site-years with a slight error rate, such as Fendt in Fendt 2020 intensive, Rottenbuch in Fendt 2015 intensive and Rottenbuch in Fendt 2015, 2016, 2017 extensive (Fig 5 c). Higher overestimations are evident for Chamau 2015 and Graswang in Fendt 2020 (Fig. 5 a, b) and extreme overestimation is evident for soils such as Rottenbuch in Fendt 2017, 2020 intensive (Fig. 5 c).

3.4 Permutation feature importance (PFI)

PFI analysis was performed on different soils in the test set to provide a comprehensive assessment of the relative importance of individual drivers. The results indicate that, for most soils, all temporal drivers were highly important for explaining N_2O dynamics (Fig. 6). This analysis revealed that soil temperature at 10 cm depth as well as fertilization were the most influential features in many soils while all other input features also showed a high relative importance (Fig. 6 a, b, e, f). In contrast, at Graswang in Fendt soils, BD and pH show a very high importance (Fig. 6 c and d), due to high variability in soil properties within the soil lysimeters which cause varying N_2O emission rates (Table 2). Compared to the soil lysimeter systems in Germany including a diverse set of soils with variability in pH and BD, the absence of variability in these properties at Chamau precludes their evaluation as potential drivers by PFI, which is reflected in the zero relative importance calculated for them by PFI as shown in Fig. 6a.

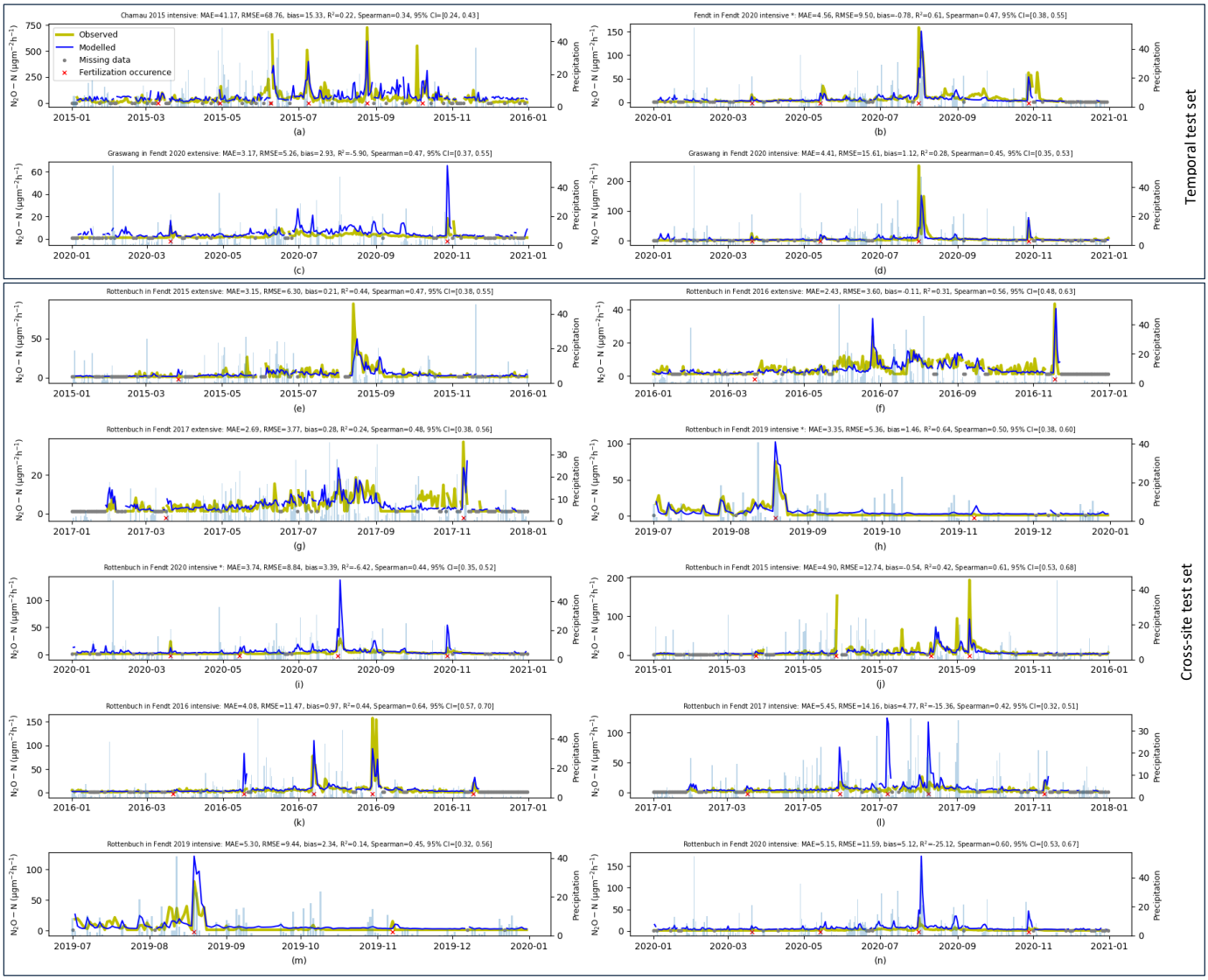


Figure 4 Time series of observed and modelled N_2O-N fluxes in different soils at different climates and years for the test set including performance metrics for individual soils in a-n panels. Panels a-d show the temporal test set and panels d-n show the cross-site test sets. Note that the axis scales differ between panels to enable visibility of the data and visualize the differences between N_2O-N emission rates at different sites in the experiment. *Originally managed extensively, and changed to intensive in 2019 and 2020.

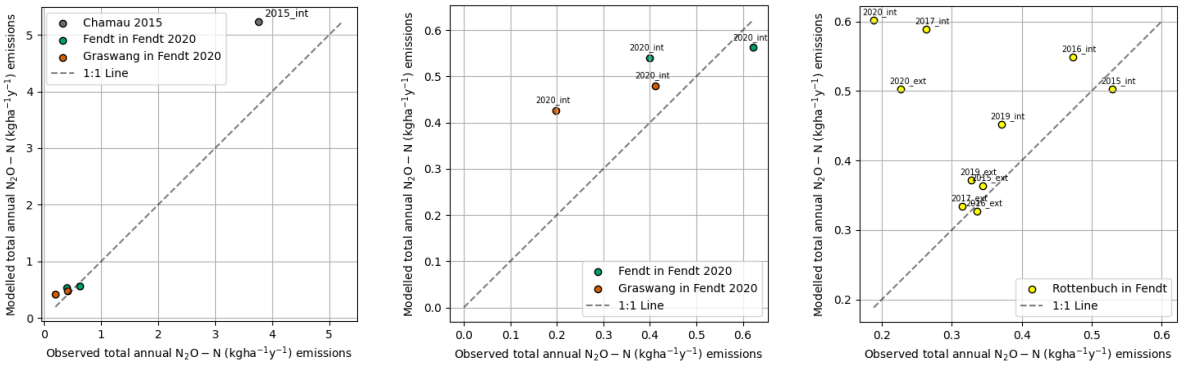


Figure 5 Cumulative observed and modelled annual N_2O-N emissions across multiple years for different soils: (a) temporal test (b) enlarged version showing only the German soils in temporal test, (c) cross-site test set (Rottenbuch in Fendt soils). Cumulative N_2O-N fluxes were calculated using only days with available observations and the corresponding simulated values, avoiding bias from gap-filling.

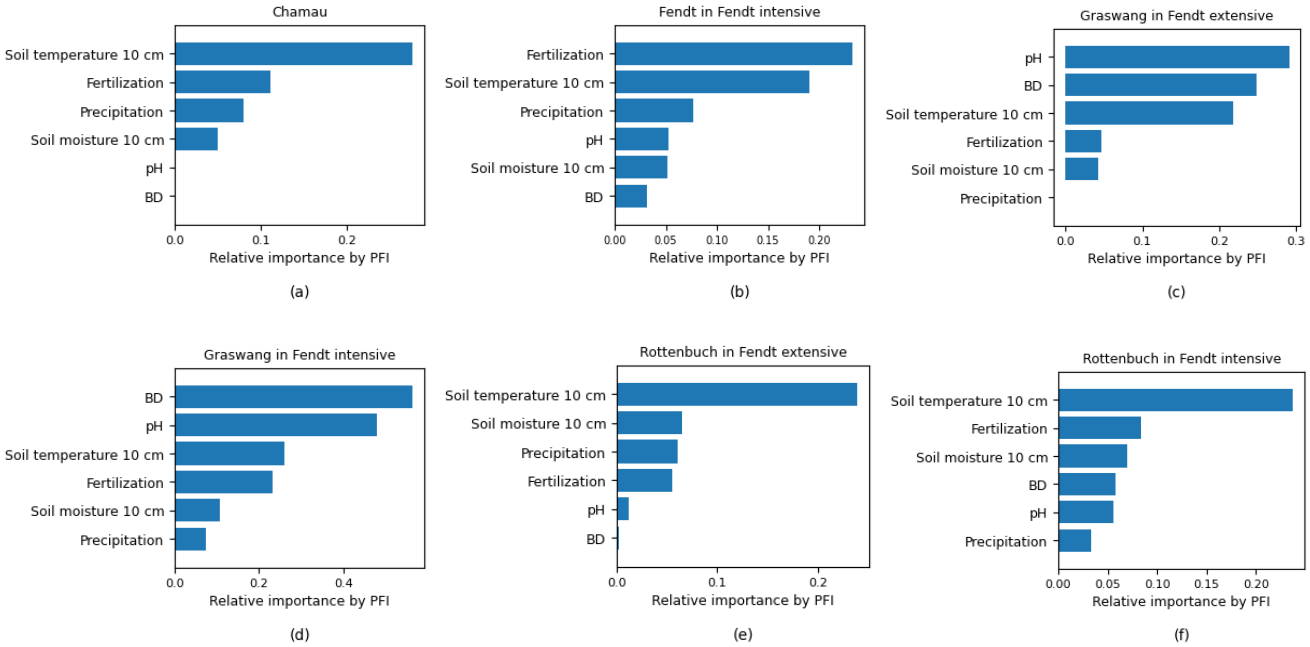


Figure 6 Relative feature importance (x-axis) of drivers (y-axis) in the test set based on MSE calculated by PFI for individual soils in panels a-f.

4 Discussion

In this study, a long short-term memory (LSTM) neural network was employed to model daily N_2O emissions from grasslands in Central Europe, based on soil properties as well as key environmental and management drivers over a 5-days window. Numerous studies have used machine-learning (ML) approaches to model N_2O emissions from soils. These applications include field-scale simulations across different land-use types (Saha et al., 2021; Hamrani et al., 2020; Jiang et al., 2023), regional or global modeling of emissions (Jiang et al., 2023; Adjuik et al., 2022), or knowledge-guided modeling approaches that integrate mechanistic process-based models with data-driven ML techniques. The latter aim to improve predictions, interpretability, and

generalization by overcoming limitations of standalone process-based or purely data-driven approaches (Liu et al., 2022; Sharma et al., 2026). A summary of these studies is given in Table 6. To further advance this field and address a key methodological gap, i.e. the limited availability of rich and diverse real-world datasets enabling spatial scaling, we expanded the modeling scope to a wide regional level for grasslands across Central Europe. In addition to temporal out-of-sample testing, the model was evaluated using a cross-site approach on an independent soil dataset to assess spatial transferability. This was achieved through a comprehensive modeling framework, the compilation of a large and diverse dataset comprising 26,500 observations, and covering a wide range of soil properties and climatic conditions (Tables 1 and 2). Furthermore, short-term environmental controls on N_2O emissions were explicitly represented using a sequential LSTM model. The resulting regional model facilitates more efficient upscaling of grassland N_2O simulations within the region, without the need for site-specific parametrization which has not previously been addressed at this scale.

Study	Gas	Data source	Data size	Input drivers	Land type	ML algorithms
Current study	N_2O	-Germany, Switzerland, Austria -10 years	28000	Soil temperature, soil moisture, precipitation, occurrence of fertilization, soil pH, soil BD	Grassland	-LSTM (deep learning)
Liu et al., 2022	N_2O	Controlled mesocosm chambers	Millions synthetic, hundreds real	Soil temperature, soil moisture, soil oxygen status, soil organic carbon, dissolved organic carbon, soil ammonium, soil nitrate, air temperature, precipitation, evapotranspiration, water balance variables, fertilizer nitrogen input, fertilization timing	Cropland	KGML
Adjuik et al., 2022	CO_2	Multiple sites from GRACENET	7863	Soil temperature, Air temperature, fertilizer amendment class, soil classification, and crop type	Cropland	-Random forest -Support vector regression -Gradient boosting (best)
Jiang et al., 2023	CH_4, N_2O	-Kunshan, China -Three years	CH_4: 552 N_2O: 558	Soil temperature, soil moisture or Irrigation volume, air temperature (maximum, minimum, and average), soil redox potential, rainfall, wind speed, and radiation	Paddy fields	-Random forest -K-nearest neighbours -Gradient boosting -Linear regression -Ensemble model of all (best)
Jiang et al., 2023	CH_4, N_2O	-Cumulative fluxes global dataset collected from 52 studies	CH_4: ~60 N_2O: ~60	Soil organic carbon, pH, texture, inorganic nitrogen fertilizer application, and non-flooded days	Paddy fields	-Ensemble model of random forest, k-nearest neighbours, gradient boosting and linear regression
Saha et al., 2021	N_2O	-2 sites in US Midwest -Six years	2246	Soil moisture, air temperature, days after fertilization, soil texture, soil carbon, precipitation, nitrogen (N) fertilizer rate or inorganic N availability in the soil	Corn rotation	-Random forest model in conjunction with a process-based model (SALUS)
Hamrani et al., 2020	CO_2, N_2O	-Quebec, Canada -Five years	126	soil temperature, soil volumetric water content, air temperature, precipitation and humidity, wind speed, surface pressure, and crop nitrogen uptake	Drainage and water table management research site	-LSTM (deep learning) -Random forest (classical ML) -Feed-forward neural networks (shallow learning)

Table 6 Review of several studies which applied different ML approaches for modelling soil N_2O emissions

4.2 Model performance and role of input drivers

The model showed variable performance across different site-years in both temporal and cross-site test sets (Table 5 and Fig. 4). Overall, performance decreased in the cross-site test compared with the temporal test.

Within the Cross-site test set, several site-year evaluations showed good predictive ability with R^2 values ranging from 0.24 to 0.64. In contrast, some site-years remained challenging, with R^2 ranging from −25 to 0.14. At the challenging site-years, substantial overestimation of N2O emissions particularly following fertilization events was noticed. These overestimations likely occurred because the model learned strong associations between fertilization events, favorable soil temperature, and soil moisture conditions from the training dataset. However, the observed emissions after such conditions remained unusually low in several instances (Fig. 4 i, l, n). This suggests that, additional controlling factors which were not represented in the model inputs have likely influenced the response. For instance, differences in substrate availability, microbial community composition, soil physical properties, or site-specific conditions also determine whether fertilization leads to substantial N_2O production. Moreover, low-emission site-years represent relatively rare conditions within the available dataset. Therefore, an expanded training set with broader range of environmental conditions and emission responses could improve the model's ability to recognize such atypical responses. Nevertheless, despite the poor R^2 values in these challenging site-years, Spearman correlation coefficients remained relatively high (0.42–0.61), indicating that the model captured the temporal dynamics of N_2O emissions even when the magnitude of individual peaks was overestimated.

In addition to the emission peaks following fertilization in the current dataset, several high peaks were caused by other climatic conditions, such as rewetting, which is an important driver of N_2O emissions (Butterbach-Bahl et al., 2013; Guo et al., 2014). For instance, at Rottenbuch in Fendt extensive in August 2015 (Fig. 4 e) as well as at Chamau intensive in October 2015 (Fig. 4 a) emission peaks were caused by 18 mm precipitation after 12 days of dry conditions and 5mm precipitation after 10 days of dry conditions, respectively. The model successfully simulated the occurrence and duration of the emission peaks at Rottenbuch but failed to capture the magnitude and exact timing (Fig. 4e). At Chamau, the model was unable to detect the occurrence of the so-called high emission. Further improvements, particularly through expansion of the training dataset, could enhance the model's ability to capture such episodic events more robustly.

Further analysis of the dataset revealed the complex interaction between the selected input variables for the model. For instance, fertilization alone did not consistently trigger immediate N2O peaks in this dataset. This aligns with the understanding that nitrogen availability must coincide with favourable environmental conditions, such as optimal soil temperature and moisture, to stimulate the microbial processes responsible for elevated fluxes (Butterbach-Bahl et al., 2013; Wang et al., 2021; Cui et al., 2024). Figure 4 b, h and j show instances of continued or delayed N_2O high fluxes extending over multiple days following fertilization events at different soils in August, which are also captured by the LSTM model.

Although the model captures daily fluctuations in many cases, inaccuracies remain in periods of very low fluxes, such as Rottenbuch in Fendt extensive in October 2017 (Fig. 4 g). During this period, changes in soil moisture, temperature and precipitation were minimal, resulting in limited variability in the driving variables and consequently reduced predicted dynamics.

During model development, we conducted a series of experiments to assess the impact of input sequence length and feature selection on model performance. Sequence lengths of 5, 10 and 20 days were tested. While longer sequences have the potential to capture long-term soil processes, their benefits were not evident in the current dataset, likely due to the dataset's limited occurrence of such events. For example, in the current dataset there are very rare cases where the high peaks after fertilizations occur later than 5 days. This suggests that the value of longer sequences may become more apparent when they are applied to larger and more diverse datasets.

4.3 Feature importance analysis

The PFI analysis conducted for individual soils in the test set, consistently demonstrates that all input features contribute significantly to model performance (Fig. 6). The substantial importance of both soil and meteorological features indicates that the model captured not only the meteorological effects, but also the influence of soil parameters on N_2O emissions. Among the predictors, soil temperature ranks the most important feature which aligns with previous

studies highlighting its critical role in regulating microbial processes (Wang et al., 2021; Butterbach-Bahl et al., 2013; Hamrani et al., 2020). Although soil temperature measurements are not always available at many monitoring sites worldwide, maximum air temperature can serve as a suitable proxy when strongly correlated with soil temperature. In the present dataset, the Pearson correlation between these variables is 0.86, suggesting that air temperature could be used with minimal impact on model performance. In the current dataset, fertilization events lead to pronounced short-term peaks in emissions, and show high importance in the PFI analysis, particularly in intensively managed soils. Figure 4 demonstrates that the model effectively captures these fertilization-induced emission peaks across soils. However, because such events are relatively infrequent in the studied soils, occurring only a few times per year, it ranks lower compared to more consistently present features (Fig. 6). This reflects a known limitation of permutation feature importance, which evaluates the expected degradation in predictive performance across the full data distribution and may therefore underestimate predictors whose effects occur mainly in rare or event-driven situations.

Soil pH and BD also appear as key drivers in the PFI analyses as shown in Fig. 6, indicating that their influence on N_2O emissions is captured by the LSTM model. This is in line with the previous studies about the influence of these two variables on soil N_2O emissions (Wang et al., 2021; Butterbach-Bahl et al., 2013). BD serves as a proxy for soil porosity and compaction, factors that strongly influence oxygen availability and thereby regulate microbial processes such as nitrification and denitrification (Li et al., 2024; Butterbach-Bahl et al., 2013). Moreover, soil pH also affects the microbial processes as near neutral soil pH values (6.6 - 8.3) are optimal for denitrification (Šimek et al., 2002), while low soil pH (<5.5) inhibits N_2O reductase, the enzyme that converts N_2O to N_2 during denitrification (Samad et al., 2016). Consequently, including soil properties such as pH and BD is essential when modelling N_2O emissions (Žurovec, O., et al., 2021; Costa Jr., C., et al., 2021).

These results indicate that the LSTM model implicitly captures the environmental controls governing the balance between nitrification and denitrification processes in soils. In particular, the combined importance of temperature, soil moisture, and oxygen-related soil properties suggests that the model learns the environmental conditions under which microbial pathways leading to N_2O production are activated.

Despite their lower rank at some soils, soil moisture and precipitation remain influential showing considerable relative importance, suggesting their added value for estimating N_2O emissions (Fig. 6). A larger dataset encompassing a wider range of emission responses to these drivers would likely further highlight their role. In addition, soil moisture often varies only slowly over extended periods, which can reduce the apparent effect detected by permutation-based importance methods. This represents a known limitation of PFI, as features with limited short-term variability may appear less important than their true mechanistic influence would suggest.

4.4 Implications and limitations

The performance of the LSTM model in capturing daily emission of the test set suggests that the model has potential for both gap-filling and predictive applications at similar alpine grassland environments, with utility for upscaling emissions across comparable regions. However, several limitations should be acknowledged.

First, a key challenge with LSTM models lies in their limited interpretability compared with non-recurrent models due to complex recurrent structure.

Second, the dataset exhibits a class imbalance, as it includes infrequent but high-magnitude emission peaks alongside prolonged periods of low emissions. This imbalance complicates the model's ability to learn the drivers of emissions, requiring more complex model implementation techniques.

Additionally, the number of available sub-daily flux measurements varied across days in Tereno and Chamau non-gap-filled datasets. The reason is that many gaps in N_2O measurements arise from quality control procedures and instrumental failures (Kiese et al., 2018; Feigenwinter et al., 2023). Such uneven data availability across different days introduces a bias into flux estimation modelling **(**Lucas-Moffat et al., 2022; Teng et al., 2020).

Finally, the dataset combined chamber and eddy covariance measurements, which differ in spatial representativeness and associated uncertainties. While this increased the diversity of conditions represented in the training data and supported model generalizability, it may also have contributed to unexplained prediction errors. Future work could investigate the influence of measurement method more explicitly. While this and several other studies have demonstrated the promise of machine learning (ML) models for simulating soil greenhouse gas (GHG) emissions, the full potential of these approaches remains underexplored. We identify three key directions for future research that could further advance the application of ML in this domain. Firstly,

expanding the dataset to encompass more geographically and climatically diverse sites with longer term interactions between the emission drivers, along with a wider range of management practices would allow for the development of more generalized N_2O emission models for grasslands across different agroecosystems. For example, grazing practices may strongly influence N_2O emission patterns but could not be systematically evaluated in this study due to limited data availability. Secondly, recent advancements in deep learning models for time series forecasting, such as transformer neural networks (Vaswani et al., 2017), could be leveraged in combination with better explainable AI tools to further enhance the progress in development of the generalized model. Transformers have demonstrated significant potential in time series modelling because of their ability to capture long-range dependencies and handle multivariate data effectively (Zerveas et al., 2021). Unlike traditional recurrent models such as LSTM, transformers process entire sequences in parallel, making them highly scalable and efficient (Vaswani et al., 2017). Finally, other versions of the machine learning models which are trained with air temperature and precipitation instead of soil temperature and moisture could be assessed using a more diverse dataset, enabling N_2O emission predictions even where direct soil measurements are missing.

5 Conclusion

This study demonstrates the effectiveness of an LSTM model in capturing temporal dynamics of daily N_2O emission data of multiple grassland sites in Central Europe, highlighting the model's suitability for predicting soil N_2O emissions. By collecting a unique dataset from a wide range of grasslands and treatments, we developed a regional model capable of accurately predicting N_2O emissions across diverse conditions in central Europe. Notably, the model's performance remained robust when tested on an independent cross-site soil dataset that was not used for training, indicating good spatial transferability. This model can therefore be used not only as a gap-filling model, but also to regionally upscale emissions in pre-alpine grasslands where the required input data are available. Future research should aim to: (1) expand the dataset to include a broader and more diverse dataset, both in terms of geographic coverage and input drivers, enhancing the model's robustness and applicability at larger scales; (2) evaluate recent advances in deep learning, such as transformer models; and (3) explore the potential of KGML approaches, to further improve prediction accuracy and model interpretability.

Appendix A

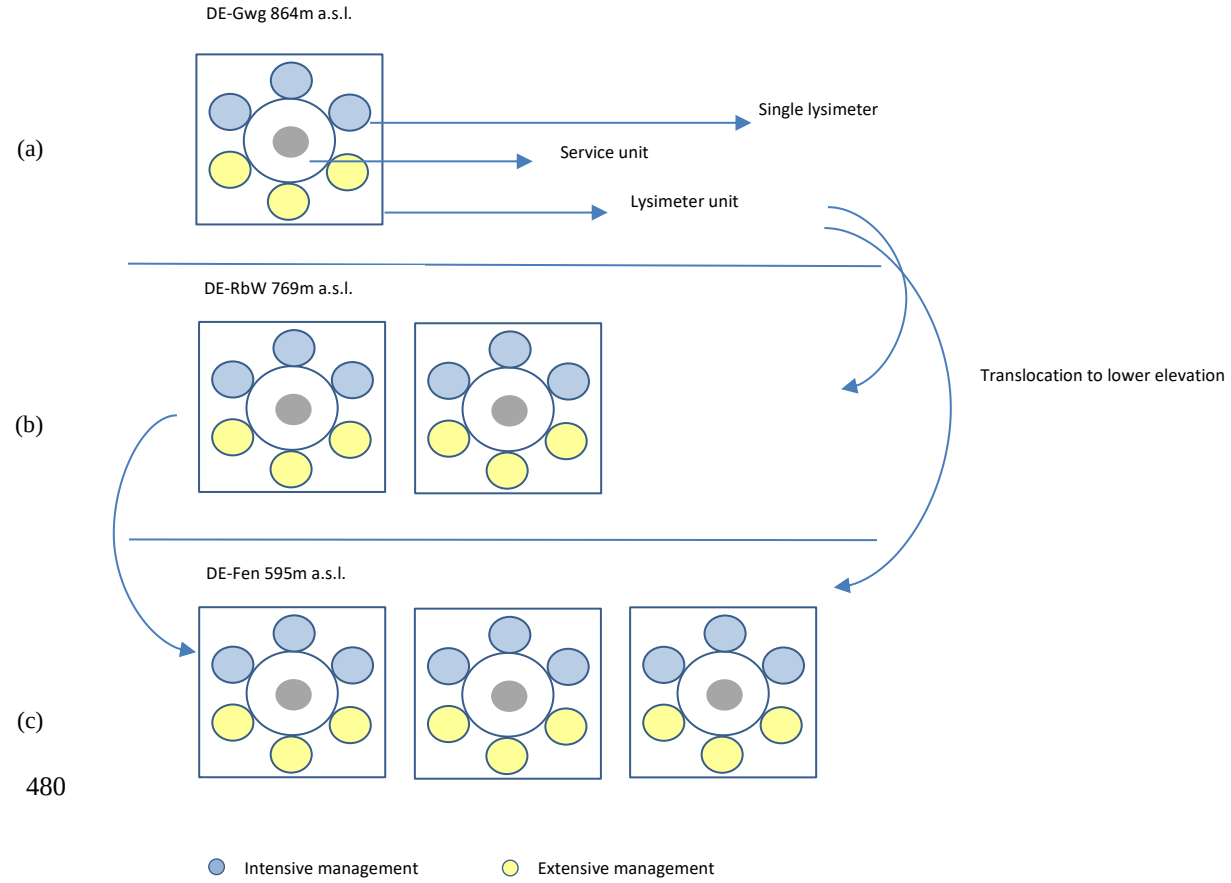


Figure A1 Scheme of lysimeter replicates originated or translocated in different grasslands, (a) Graswang (DE-Gwg), (b) Rottenbuch (DE-RbW), and (c) Fendt (DE-Fen) within the TERENO Pre-Alpine Observatory in Germany (Kiese et al., 2018).

Soil	Replicate-Management	Mean	Std	Min	Max
Neustift		20.87	6.08	7.94	37.89
Chamau		43.84	9.02	16.93	54.98
Fendt in Fendt	L1-Intensive	42.04	6.78	17.85	54.97
Fendt in Fendt	L2-Intensive	40.34	6.24	18.22	51.47
Fendt in Fendt	L3-Intensive	41.64	6.99	17.85	54.58
Fendt in Fendt	L1-Extensive (2014-2018), Intensive (2019-2020)	42.96	6.67	17.85	53.90
Fendt in Fendt	L2-Extensive (2014-2018), Intensive (2019-2020)	40.09	7.44	16.09	53.90
Fendt in Fendt	L3-Extensive (2014-2018), Intensive (2019-2020)	43.51	7.86	17.85	54.59
Graswang in Fendt	L1-Intensive	44.44	7.22	17.28	56.14
Graswang in Fendt	L2-Intensive	43.54	6.92	19.37	53.39
Graswang in Fendt	L3-Intensive	46.25	5.90	21.01	53.56
Graswang in Fendt	L1-Extensive	45.76	7.40	19.60	57.05
Graswang in Fendt	L2-Extensive	43.64	7.04	14.90	57.65
Graswang in Fendt	L3-Extensive	40.78	6.74	19.60	54.34
Rottenbuch in Fendt	L1-Intensive	41.59	6.34	16.90	51.03
Rottenbuch in Fendt	L2-Intensive	38.97	6.25	15.56	51.70
Rottenbuch in Fendt	L3-Intensive	43.36	6.58	17.07	52.79
Rottenbuch in Fendt	L1-Extensive (2014-2018), Intensive (2019-2020)	41.68	5.55	17.68	50.20
Rottenbuch in Fendt	L2-Extensive (2014-2018), Intensive (2019-2020)	42.42	7.13	15.94	50.67
Rottenbuch in Fendt	L3-Extensive (2014-2018), Intensive (2019-2020)	44.05	7.26	17.63	53.69
Graswang in Graswang	L1-Intensive	47.15	7.80	13.38	58.15
Graswang in Graswang	L2-Intensive	46.73	5.41	18.37	53.28
Graswang in Graswang	L3-Intensive	48.05	6.20	17.28	55.53
Graswang in Graswang	L1-Extensive	47.89	5.78	16.98	55.85
Graswang in Graswang	L2-Extensive	49.08	6.42	19.37	57.94
Graswang in Graswang	L3-Extensive	47.89	5.63	21.19	55.85

(a)

Soil	Replicate-Management	Mean	Std	Min	Max
Neustift		9.36	6.76	0.31	19.41
Chamau		12.09	6.46	1.18	25.08
Fendt in Fendt	L1-Intensive	10.80	6.82	0.49	24.40
Fendt in Fendt	L2-Intensive	10.55	6.91	0.25	24.38
Fendt in Fendt	L3-Intensive	11.02	6.86	0.40	24.42
Fendt in Fendt	L1-Extensive (2014-2018), Intensive (2019-2020)	10.87	6.98	0.23	25.09
Fendt in Fendt	L2-Extensive (2014-2018), Intensive (2019-2020)	10.88	6.74	0.56	24.31
Fendt in Fendt	L3-Extensive (2014-2018), Intensive (2019-2020)	10.92	6.91	-0.77	24.82
Graswang in Fendt	L1-Intensive	10.53	6.87	0.63	24.74
Graswang in Fendt	L2-Intensive	10.73	6.74	0.45	24.79
Graswang in Fendt	L3-Intensive	10.86	6.80	-0.44	24.59
Graswang in Fendt	L1-Extensive	10.79	6.86	0.14	24.31
Graswang in Fendt	L2-Extensive	11.42	6.66	1.02	25.14
Graswang in Fendt	L3-Extensive	10.68	6.80	-1.28	24.38
Rottenbuch in Fendt	L1-Intensive	10.83	6.72	-0.36	25.14
Rottenbuch in Fendt	L2-Intensive	11.07	6.97	-0.22	24.96
Rottenbuch in Fendt	L3-Intensive	10.75	6.76	-0.30	24.58
Rottenbuch in Fendt	L1-Extensive (2014-2018), Intensive (2019-2020)	10.48	6.59	-0.26	24.92
Rottenbuch in Fendt	L2-Extensive (2014-2018), Intensive (2019-2020)	11.16	6.56	0.74	24.40
Rottenbuch in Fendt	L3-Extensive (2014-2018), Intensive (2019-2020)	10.82	6.68	-0.36	24.48
Graswang in Graswang	L1-Intensive	9.40	6.93	-1.35	23.06
Graswang in Graswang	L2-Intensive	8.89	6.70	-1.35	22.20
Graswang in Graswang	L3-Intensive	9.31	6.80	-1.21	22.32
Graswang in Graswang	L1-Extensive	9.34	6.64	-0.71	22.64
Graswang in Graswang	L2-Extensive	9.28	6.50	-1.06	22.18
Graswang in Graswang	L3-Extensive	9.45	6.62	-1.43	22.73

(b)

Soil	Replicate-Management	Mean	Std	Min	Max
Neustift		12.54	10.41	0.11	48.86
Chamau		70.11	114.03	-739.12	1155.83
Fendt in Fendt	L1-Intensive	5.84	13.56	-25.01	248.33
Fendt in Fendt	L2-Intensive	4.52	7.79	-26.17	160.07
Fendt in Fendt	L3-Intensive	11.44	30.64	-25.68	707.50
Fendt in Fendt	L1-Extensive (2014-2018), Intensive (2019-2020)	6.27	19.81	-30.72	357.82
Fendt in Fendt	L2-Extensive (2014-2018), Intensive (2019-2020)	6.13	12.79	-15.00	192.32
Fendt in Fendt	L3-Extensive (2014-2018), Intensive (2019-2020)	14.97	48.18	-24.85	750.38
Graswang in Fendt	L1-Intensive	7.16	35.14	-25.57	960.49
Graswang in Fendt	L2-Intensive	5.77	14.12	-26.52	288.06
Graswang in Fendt	L3-Intensive	6.10	16.39	-20.97	326.71
Graswang in Fendt	L1-Extensive	6.09	17.05	-28.28	428.09
Graswang in Fendt	L2-Extensive	4.88	8.21	-26.73	164.57
Graswang in Fendt	L3-Extensive	3.52	5.80	-27.29	79.14
Rottenbuch in Fendt	L1-Intensive	4.30	10.84	-30.25	226.26
Rottenbuch in Fendt	L2-Intensive	4.40	13.25	-22.33	275.85
Rottenbuch in Fendt	L3-Intensive	6.09	16.04	-10.27	323.24
Rottenbuch in Fendt	L1-Extensive (2014-2018), Intensive (2019-2020)	2.84	4.03	-15.93	52.58
Rottenbuch in Fendt	L2-Extensive (2014-2018), Intensive (2019-2020)	4.09	6.15	-14.12	81.62
Rottenbuch in Fendt	L3-Extensive (2014-2018), Intensive (2019-2020)	5.68	10.92	-13.35	132.58
Graswang in Graswang	L1-Intensive	14.14	36.82	0.22	513.37
Graswang in Graswang	L2-Intensive	11.78	26.53	1.19	230.11
Graswang in Graswang	L3-Intensive	9.01	22.17	-0.48	283.05
Graswang in Graswang	L1-Extensive	11.37	21.43	1.14	222.33
Graswang in Graswang	L2-Extensive	8.33	15.67	0.46	103.26
Graswang in Graswang	L3-Extensive	4.90	5.96	1.12	49.66

(c)

Table A1 Statistics of the soils in different climates with respect to (a) soil moisture at 0–10 cm depth, (b) soil temperature at 0–10 cm depth, and (c) daily N_2O emissions.

Code availability

The code to the implemented LSTM model architecture in python is available under: https://doi.org/10.5281/zenodo.20842387.

Data availability

The datasets analyzed during the current study were sourced from existing publications.

Author contributions

SP conceptualization, investigation, data collection and analysis, model implementation, visualization, writing -original draft preparation, -review and editing. KF conceptualization, writing -review and editing, supervision. BW: data provision, writing -review and editing. RK: data provision, writing -review and editing, supervision. CS conceptualization, writing -review and editing, supervision, funding acquisition, project administration.

Competing interest

The authors declare that they have no competing interests.

Acknowledgment

We thank George Wohlfahrt (University of Innsbruck), Lukas Hörtnagl, Nina Buchmann and Iris Feigenwinter (ETH Zurich) for supporting us in compiling the datasets from Neustift Austria (Funded by EU FP 5 project CARBOMONT (EVK2-CT2001-00125), the Austrian National Science Fund (P17560), and the Tyrolean Science Fund (Uni-404/33)) and Chamau Switzerland (funded by EU project SUPER-G (grant number 774124), the SNF projects GrassGas (200021-105949) and M4P (40FA40_154245)).

References

- Abdalla, M., Jones, M., Yeluripati, J., Smith, P., Burke, J., and Williams, M.: Testing DayCent and DNDC model simulations of N_2O fluxes and assessing the impacts of climate change on the gas flux and biomass production from a humid pasture, Atmos. Environ., 44(25), 2961–2970, https://doi.org/10.1016/j.atmosenv.2010.05.018, 2010.
- Adjuik, T. A., and Davis, S. C.: Machine learning approach to simulate soil CO_2 fluxes under cropping systems, Agronomy, 12, 197, https://doi.org/10.3390/agronomy12010197, 2022.
- Bahdanau, D., Cho, K., and Bengio, Y.: Neural machine translation by jointly learning to align and translate, in: Proc. 3rd Int. Conf. Learn. Represent. (ICLR), https://arxiv.org/abs/1409.0473, 2015.
- Bigaignon, L., Fieuzal, R., Delon, C., and Tallec, T.: Combination of two methodologies, artificial neural network and linear interpolation, to gap-fill daily nitrous oxide flux measurements, Agric. For. Meteorol., 291, 108037, https://doi.org/10.1016/j.agrformet.2020.108037, 2020.
- Bonett, D. G., and Wright, T. A.: Sample size requirements for estimating Pearson, Kendall and Spearman correlations, Psychometrika, 65, 23–28, https://doi.org/10.1007/BF02294183, 2000.
- Box, G. E. P., and Cox, D. R.: An analysis of transformations, J. R. Stat. Soc. B, 26, 211–243, https://doi.org/10.1111/j.2517-6161.1964.tb00553.x, 1964.
- Breiman, L.: Random forests, Mach. Learn., 45(1), 5–32, https://doi.org/10.1023/A:1010933404324, 2001.
- Brilli, L., Bechini, L., Bindi, M., Carozzi, M., Cavalli, D., Conant, R., Dorich, C. D., Doro, L., Ehrhardt, F., Farina, R., Ferrise, R., Fitton, N., Francaviglia, R., Grace, P., Iocola, I., Klumpp, K., Léonard, J., Martin, R., Massad, R. S., Recous, S., Seddaiu, G., Sharp, J., Smith, P., Smith, W. N., Soussana, J.-F., and Bellocchi, G.: Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C and N fluxes, Sci. Total Environ., 598, 445–470, https://doi.org/10.1016/j.scitotenv.2017.03.208, 2017.
- Butterbach-Bahl, K., Baggs, E. M., Dannenmann, M., Kiese, R., and Zechmeister-Boltenstern, S.: Nitrous oxide emissions from soils: How well do we understand the processes and their controls?, Philos. Trans. R. Soc. B, 368, 20130122, https://doi.org/10.1098/rstb.2013.0122, 2013.
- Chataut, G., Bhatta, B., Joshi, D., Subedi, K., and Kafle, K.: Greenhouse gases emission from agricultural soil: A review, J. Agric. Food Res., 11, 100533, https://doi.org/10.1016/j.jafr.2023.100533, 2023.
- Chhajer, P., Shah, M., and Kshirsagar, A.: The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction, Decis. Anal. J., 2, 100015, https://doi.org/10.1016/j.dajour.2021.100015, 2022.
- Chollet, F., et al.: Keras, https://keras.io, 2015.
- Costa Jr., C., Li, C., Cerri, C. E. P., and Cerri, C. C.: Modeling nitrous oxide emissions from large-scale intensive cropping systems in the Southern Amazon, Front. Sustain. Food Syst., 5, 701416, https://doi.org/10.3389/fsufs.2021.701416, 2021.
- Cui, X., Bo, Y., Adalibieke, W., Winiwarter, W., Zhang, X., Davidson, E. A., Sun, Z., Tian, H., Smith, P., and Zhou, F.: The global potential for mitigating nitrous oxide emissions from croplands, One Earth, 7(3), 401–420, https://doi.org/10.1016/j.oneear.2024.01.005, 2024.
- Denk, T. R. A., Kraus, D., Kiese, R., Butterbach-Bahl, K., and Wolf, B.: Constraining N cycling in the ecosystem model LandscapeDNDC with the stable isotope model SIMONE, Ecology, 100, e02675, https://doi.org/10.1002/ecy.2675, 2019.
- Ehrhardt, F., Soussana, J.-F., Bellocchi, G., Grace, P., McAuliffe, R., Recous, S., Sandor, R., Smith, P., Snow, V., de Antoni Migliorati, M., Basso, B., Bhatia, A., Brilli, L., Doltra, J., Dorich, C. D., Doro, L., Fitton, N., Giacomini, S. J., Grant, B., Harrison, M. T., Jones, S. K., Kirschbaum, M. U. F., Klumpp, K., Laville, P., Léonard, J., Liebig, M., Lieffering, M., Martin, R., Massad, R. S., Meier, E., Merbold, L., Moore, A. D., Myrgiotis, V., Newton, P., Pattey, E., Rolinski, S., Sharp, J., Smith, W. N., Wu, L., and Zhang, Q.: Assessing uncertainties in crop and pasture ensemble model simulations of productivity and N_2O emissions, Glob. Change Biol., 24, e603–e616, https://doi.org/10.1111/gcb.13965, 2018.
- Feigenwinter, I., Hörtnagl, L., and Buchmann, N.: N_2O and CH_4 fluxes from intensively managed grassland: The importance of biological and environmental drivers vs. management, Sci. Total Environ., 903, 166389, https://doi.org/10.1016/j.scitotenv.2023.166389, 2023.

- Fitton, N., Datta, A., Hastings, A., Kuhnert, M., Topp, C. F. E., Cloy, J. M., Rees, R. M., Cardenas, L. M., Williams, J. R., Smith, K., Chadwick, D., and Smith, P.: The challenge of modelling nitrogen management at the field scale: simulation and sensitivity analysis of N_2O fluxes across nine experimental sites using DailyDayCent, Environ. Res. Lett., 9, 095003, https://doi.org/10.1088/1748-9326/9/9/095003, 2014.
- Fuchs, K., Hörtnagl, L., Buchmann, N., Eugster, W., Snow, V., and Merbold, L.: Management matters: Testing a mitigation strategy for nitrous oxide emissions using legumes on intensively managed grassland, Biogeosciences, 15, 5519–5534, https://doi.org/10.5194/bg-15-5519-2018, 2018.
- Gilgen, A. K., and Buchmann, N.: Response of temperate grasslands at different altitudes to simulated summer drought differed but scaled with annual precipitation, Biogeosciences, 6, 2525–2539, https://doi.org/10.5194/bg-6-2525-2009, 2009.
- Giltrap, D. L., Li, C., and Saggar, S.: DNDC: A process-based model of greenhouse gas fluxes from agricultural soils, Agric. Ecosyst. Environ., 136, 292–300, https://doi.org/10.1016/j.agee.2009.07.006, 2010.
- Goodfellow, I., Bengio, Y., and Courville, A.: Deep learning, MIT Press, 2016.
- Guo, X., Drury, C. F., Yang, X., Reynolds, W. D., and Fan, R.: The extent of soil drying and rewetting affects nitrous oxide emissions, denitrification, and nitrogen mineralization, Soil Sci. Soc. Am. J., 78(1), 194–204, https://doi.org/10.2136/sssaj2013.06.0219, 2014.
- Haas, E., Klatt, J., Fröhlich, A., Kraft, P., Werner, C., Kiese, R., Grote, R., Breuer, L., and Butterbach-Bahl, K.: LandscapeDNDC: A process model for simulation of biosphere-atmosphere-hydrosphere exchange processes at site and regional scale, Landscape Ecology, 28, 615–636, https://doi.org/10.1007/s10980-012-9772-x, 2013.
- Hamrani, A., Akbarzadeh, A., and Madramootoo, C. A.: Machine learning for predicting greenhouse gas emissions from agricultural soils, Sci. Total Environ., 741, 140338, https://doi.org/10.1016/j.scitotenv.2020.140338, 2020.
- Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: Data mining, inference, and prediction (2nd ed.), Springer, 2009.
- Hunter, J. D.: Matplotlib: A 2D graphics environment, Comput. Sci. Eng., 9(3), 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.
- Hyndman, R. J., and Athanasopoulos, G.: Forecasting: Principles and practice (3rd ed.), OTexts, https://otexts.com/fpp3/, 2021.
- Hodson, T. O.: Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not, Geoscientific Model Development, 15, 5481–5487, https://doi.org/10.5194/gmd-15-5481-2022, 2022.
- Hörtnagl, L., Barthel, M., Buchmann, N., Eugster, W., Butterbach-Bahl, K., Díaz-Pinés, E., Zeeman, M., Klumpp, K., Kiese, R., Bahn, M., Hammerle, A., Lu, H., Ladreiter-Knauss, T., Burri, S., and Merbold, L.: Greenhouse gas fluxes over managed grasslands in Central Europe, Glob. Change Biol., 24, 1843–1872, https://doi.org/10.1111/gcb.14079, 2018.
- Hörtnagl, L., and Wohlfahrt, G.: Methane and nitrous oxide exchange over a managed hay meadow, Biogeosciences, 11, 7219–7236, https://doi.org/10.5194/bg-11-7219-2014, 2014.
- Hörtnagl, L., Feigenwinter, I., Wang, Y., and others / Project leader: Nina Buchmann: Eddy covariance ecosystem fluxes, meteorological data and detailed management information for the intensively managed grassland site Chamau in Switzerland, collected between 2005 and 2024, Dataset, ETH Zürich, https://doi.org/10.3929/ethz-b-000747025, 2025.
- Jiang, Z., Yang, S., Smith, P., and Pang, Q.: Ensemble machine learning for modeling greenhouse gas emissions at different time scales from irrigated paddy fields, *Field Crops Research*, **292**, 108821, https://doi.org/10.1016/j.fcr.2023.108821, 2023.
- Khalil, M. I., Abdalla, M., Lanigan, G., Osborne, B., and Müller, C.: Evaluation of parametric limitations in simulating greenhouse gas fluxes from Irish arable soils using three process-based models, Agric. Sci., 07(08), 503–520, https://doi.org/10.4236/as.2016.78051, 2016.
- Khoshvaght, H., Permala, R. R., Razmjou, A., and Khiadani, M.: A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction, J. Environ. Chem. Eng., 13(6), 119675, https://doi.org/10.1016/j.jece.2025.119675, 2025.
- Kiese, R., Fersch, B., Baessler, C., Brosy, C., Butterbach-Bahl, K., Chwala, C., Dannenmann, M., Fu, J., Gasche, R., Grote, R., Jahn, C., Klatt, J., Kunstmann, H., Mauder, M., Rödiger, T., Smiatek, G., Soltani, M., Steinbrecher, R., Völksch, I., Werhahn, J., Wolf, B., Zeeman, M., and

Schmid, H. P.: The TERENO Pre-Alpine Observatory: Integrating meteorological, hydrological, and biogeochemical measurements and modelling, Vadose Zone J., 17, 1–17, https://doi.org/10.2136/vzj2018.03.0060, 2018.

- Kopittke, P. M., Dalal, R. C., McKenna, B. A., Smith, P., Wang, P., Weng, Z., van der Bom, F. J. T., and Menzies, N. W.: Soil is a major contributor to global greenhouse gas emissions and climate change, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2024-1782, 2024.
- Le, X.-H., Ho, H. V., Lee, G., and Jung, S.: Application of Long Short-Term Memory (LSTM) Neural Network for flood forecasting, Water, 11(7), 1387, https://doi.org/10.3390/w11071387, 2019.
- Li, C., Frolking, S., and Frolking, T. A.: A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity, J. Geophys. Res.-Atmos., 97, 9759–9776, https://doi.org/10.1029/92JD00509, 1992.
- Li, D., Li, Y., Yao, S., Zhou, H., Huang, S., Peng, X., and Meng, Y.: Dynamics of nitrogen mineralization and nitrogen cycling functional genes in response to soil pore size distribution, Eur. J. Soil Biol., 123, 103692, https://doi.org/10.1016/j.ejsobi.2024.103692, 2024.
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A.: Hyperband: A novel bandit-based approach to hyperparameter optimization, J. Mach. Learn. Res., 18, 1–52, https://arxiv.org/abs/1603.06560, 2016.
- Li, Y., Chen, D., Zhang, Y., Edis, R., and Ding, H.: Comparison of three modeling approaches for simulating denitrification and nitrous oxide emissions from loam-textured arable soils: Modeling denitrification and nitrous oxide, Global Biogeochemical Cycles, 19, GB3002, https://doi.org/10.1029/2004GB002392, 2005.
- Lin, C., Yeap, T., and Kiringa, I.: Stacked bidirectional LSTM for predicting emission of nitrous oxide, in: Proc. Can. Conf. Artif. Intell., https://doi.org/10.21428/594757db.daad1be1, 2022.
- Liu, L., Xu, S., Tang, J., Guan, K., Griffis, T. J., Erickson, M. D., Frie, A. L., Jia, X., Kim, T., Miller, L. T., Peng, B., Wu, S., Yang, Y., Zhou, W., Kumar, V., and Jin, Z.: KGML-ag: A modeling framework of knowledge-guided machine learning to simulate agroecosystems: A case study of estimating N_2O emission using data from mesocosm experiments, Geosci. Model Dev., 15, 2839–2858, https://doi.org/10.5194/gmd-15-2839-2022, 2022.
- Lucas-Moffat, A. M., Schrader, F., Herbst, M., and Brümmer, C.: Multiple gap-filling for eddy covariance datasets, Agric. For. Meteorol., 325, 109114, https://doi.org/10.1016/j.agrformet.2022.109114, 2022.
- Mauder, M., Cuntz, M., Drüe, C., Graf, A., Rebmann, C., Schmid, H. P., Schmidt, M., and Steinbrecher, R.: A strategy for quality and uncertainty assessment of long-term eddy-covariance measurements, Agric. For. Meteorol., 169, 122–135, https://doi.org/10.1016/j.agrformet.2012.09.006, 2013.
- Merbold, L., Eugster, W., Stieger, J., Zahniser, M., Nelson, D., and Buchmann, N.: Greenhouse gas budget (CO_2, CH_4 and N_2O) of intensively managed grassland following restoration, Glob. Change Biol., 20(6), 1913–1928, https://doi.org/10.1111/gcb.12518, 2014.
- Montgomery, D. C., Peck, E. A., and Vining, G. G.: Introduction to linear regression analysis (6th ed.), Wiley, 2021.
- Parton, W. J., Hartman, M., Ojima, D., and Schimel, D.: DAYCENT and its land surface submodel: Description and testing, Glob. Planet. Change, 19, 35–48, https://doi.org/10.1016/S0921-8181(98)00040-X, 1998.
- Parton, W. J., Mosier, A. R., Ojima, D. S., Valentine, D. W., Schimel, D. S., Weier, K., and Kulmala, A. E.: Generalized model for NOx and N_2O emissions from soils, Global Biogeochemical Cycles, 10(3), 401–412, https://doi.org/10.1029/96GB01455, 2001.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825–2830, https://doi.org/10.48550/arXiv.1201.0490, 2011.
- Prion, S., and Haerling, K. A.: Making Sense of Methods and Measurement: Spearman-Rho Ranked-Order Correlation Coefficient, Clin. Simul. Nurs., 10(10), 587–588, https://doi.org/10.1016/j.ecns.2014.07.005, 2014.
- Pütz, T., Kiese, R., Wollschläger, U., Groh, J., Rupp, H., Zacharias, S., Priesack, E., Gerke, H. H., Gasche, R., Bens, O., Borg, E., Baessler, C., Kaiser, K., Herbrich, M., Munch, J.-C., Sommer, M., Vogel, H.-J., Vanderborght, J., and Vereecken, H.: TERENO-SOILCan: a lysimeter-

network in Germany observing soil processes and plant diversity influenced by climate change, *Environmental Earth Sciences*, **75**, 1242, https://doi.org/10.1007/s12665-016-6031-5, 2016.

- Roth, K.: Bodenkartierung und GIS-basierte Kohlenstoffinventur von Graslandböden, diploma thesis, University of Zurich, Zurich, Switzerland, 2006.
- Ruscio, J.: Constructing confidence intervals for Spearman's rank correlation with ordinal data: A simulation study comparing analytic and bootstrap methods, J. Mod. Appl. Stat. Methods, 7, 416–434, https://doi.org/10.22237/jmasm/1225513620, 2008.
- Saggar, S., Jha, N., Deslippe, J., Bolan, N. S., Luo, J., Giltrap, D. L., Kim, D.-G., Zaman, M., and Tillman, R. W.: Denitrification and N_2O:N_2 production in temperate grasslands: Processes, measurements, modelling and mitigating negative impacts, Sci. Total Environ., 465, 173–195, https://doi.org/10.1016/j.scitotenv.2013.01.050, 2013.
- Saha, D., Basso, B., and Robertson, G. P.: Machine learning improves predictions of agricultural nitrous oxide (N_2O) emissions from intensively managed cropping systems, Environ. Res. Lett., 16, https://doi.org/10.1088/1748-9326/abd2f3, 2021.
- Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T.: DeepAR: probabilistic forecasting with autoregressive recurrent networks, Int. J. Forecasting, 36(3), 1181–1191, https://doi.org/10.1016/j.ijforecast.2019.07.001, 2020.
- Samad, M. S., Bakken, L. R., Nadeem, S., Clough, T. J., de Klein, C. A. M., Richards, K. G., Lanigan, G., and Morales, S. E.: High-resolution denitrification kinetics in pasture soils link N_2O emissions to pH, and denitrification to C mineralization, *PLOS ONE*, **11**, e0151713, https://doi.org/10.1371/journal.pone.0151713, 2016.
- Sharma, P., Basso, B., Manuraj, A., Murillo, M. S., Millar, N., Tadiello, T., Sharma, M., Delandmeter, M., and Robertson, G. P.: Coupled machine learning–ecosystem ensemble models substantially improve predictions of nitrous oxide (N_2O) fluxes from US croplands, Proc. Natl. Acad. Sci. U.S.A., 123, e2524808123, https://doi.org/10.1073/pnas.2524808123, 2026.
- Smith, P., Martino, D., Cai, Z., Gwary, D., Janzen, H., Kumar, P., McCarl, B., Ogle, S., O'Mara, F., Rice, C., Scholes, B., Sirotenko, O., Howden, M., McAllister, T., Pan, G., Romanenkov, V., Schneider, U., Towprayoon, S., Wattenbach, M., and Smith, J.: Agriculture, in: *Climate Change 2007: Mitigation of Climate Change. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by: Metz, B., Davidson, O. R., Bosch, P. R., Dave, R., and Meyer, L. A., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 497–540, 2007.
- Smith, P., Keesstra, S. D., Silver, W. L., and Adhya, T. K.: The role of soils in delivering Nature's Contributions to People, *Philosophical Transactions of the Royal Society B: Biological Sciences*, **376**, 20200169, https://doi.org/10.1098/rstb.2020.0169, 2021.
- Smyl, S.: A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting, International Journal of Forecasting, 36(1), 75–85, https://doi.org/10.1016/j.ijforecast.2019.03.017, 2020.
- Soltani, M., Mauder, M., Laux, P., and Kunstmann, H.: Turbulent flux variability and energy balance closure in the TERENO prealpine observatory: A hydrometeorological data analysis, Theor. Appl. Climatol., 133, 937–956, https://doi.org/10.1007/s00704-017-2235-z, 2018.
- Spearman, C.: The proof and measurement of association between two things, Am. J. Psychol., 15, 72–101, https://doi.org/10.2307/1412159, 1904.
- Stehfest, E., and Müller, C.: Simulation of N_2O emissions from a urine-affected pasture in New Zealand with the ecosystem model DayCent: Simulation of N_2O from grassland, J. Geophys. Res., 109, D03109, https://doi.org/10.1029/2003JD004261, 2004.
- Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M. (Eds.): *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp., 2013.
- Sánchez-Martín, L., Vallejo, A., Dick, J., and Skiba, U. M.: The influence of soluble carbon and fertilizer nitrogen on nitric oxide and nitrous oxide emissions from two contrasting agricultural soils, Soil Biol. Biochem., 40, 142–151, https://doi.org/10.1016/j.soilbio.2007.07.016, 2008.
- Teng, D., He, X., Wang, J., Wang, J., and Lv, G.: Uncertainty in gap filling and estimating the annual sum of carbon dioxide exchange for the desert Tugai forest, Ebinur Lake Basin, Northwest China, PeerJ, 8, e8530, https://doi.org/10.7717/peerj.8530, 2020.

- Thorburn, P. J., Biggs, J. S., Collins, K., and Probert, M. E.: Using the APSIM model to estimate nitrous oxide emissions from diverse Australian sugarcane production systems, Agric. Ecosyst. Environ., 136(3), 343–350, https://doi.org/10.1016/j.agee.2009.12.014, 2010.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, Adv. Neural Inf. Process. Syst., 30, 5998–6008, https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf, 2017.
- Wang, C., Amon, B., Schulz, K., and Mehdi, B.: Factors that influence nitrous oxide emissions from agricultural soils as well as their representation in simulation models: A review, Agronomy, 11, 770, https://doi.org/10.3390/agronomy11040770, 2021.
- West, R. M.: Best practice in statistics: the use of log transformation, Ann. Clin. Biochem., 59(3), 162–165, https://doi.org/10.1177/00045632211050531, 2022.
- Willmott, C. J., and Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Clim. Res., 30(1), 79–82, https://doi.org/10.3354/cr030079, 2005.
- Wohlfahrt, G., Hammerle, A., Haslwanter, A., Bahn, M., Tappeiner, U., and Cernusca, A.: Seasonal and inter-annual variability of the net ecosystem CO_2 exchange of a temperate mountain grassland: Effects of weather and management, J. Geophys. Res., 113, D08110, https://doi.org/10.1029/2007JD009286, 2008.
- Wohlfahrt, G., Hammerle, A., and Hörtnagl, L.: FLUXNET2015 AT-Neu Neustift, Dataset, https://doi.org/10.18140/FLX/1440121, 2012.
- Xing, H., Wang, E., Smith, C. J., Rolston, D., and Yu, Q.: Modelling nitrous oxide and carbon dioxide emission from soil in an incubation experiment, Geoderma, 167–168, 328–339, https://doi.org/10.1016/j.geoderma.2011.07.003, 2011.
- Zhang, J., Zhu, Y., Zhang, X., Ye, M., and Yang, J.: Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas, J. Hydrol., 561, 918–929, https://doi.org/10.1016/j.jhydrol.2018.04.065, 2018.
- Zimmermann, J., Carolan, R., Forrestal, P., Harty, M., Lanigan, G., Richards, K. G., Roche, L., Whitfield, M. G., and Jones, M. B.: Assessing the performance of three frequently used biogeochemical models when simulating N_2O emissions from a range of soil types and fertiliser treatments, Geoderma, 331, 53–69, https://doi.org/10.1016/j.geoderma.2018.06.004, 2018.
- Šimek, M., Jíšová, L., and Hopkins, D. W.: What is the so-called optimum pH for denitrification in soil?, Soil Biol. Biochem., 34(9), 1227–1234, https://doi.org/10.1016/S0038-0717(02)00059-7, 2002.
- Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., and Eickhoff, C.: A transformer-based framework for multivariate time series representation learning, Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2114–2124, https://doi.org/10.1145/3447548.3467401, 2021.
- Žurovec, O., Wall, D. P., Brennan, F. P., Krol, D. J., Forrestal, P. J., and Richards, K. G.: Increasing soil pH reduces fertiliser-derived N_2O emissions in intensively managed temperate grassland, Agric. Ecosyst. Environ., 311, 107319, https://doi.org/10.1016/j.agee.2021.107319, 2021.