Author/s:
Pan, B;Lam, SK;Wang, E;Mosier, A;Chen, D

Title:
New approach for predicting nitrification and its fraction of N2O emissions in global terrestrial ecosystems

Date:
2021-03-01

Citation:
Pan, B., Lam, S. K., Wang, E., Mosier, A. & Chen, D. (2021). New approach for predicting nitrification and its fraction of N2O emissions in global terrestrial ecosystems. Environmental Research Letters, 16 (3), https://doi.org/10.1088/1748-9326/abe4f5.

Persistent Link:
https://hdl.handle.net/11343/290193

ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

New approach for predicting nitrification and its fraction of N_2O emissions in global terrestrial ecosystems

To cite this article: Baobao Pan *et al* 2021 *Environ. Res. Lett.* **16** 034053

IOP Publishing Environ. Res. Lett. 16 (2021) 034053 https://doi.org/10.1088/1748-9326/abe4f5

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
11 September 2020

REVISED
2 February 2021

ACCEPTED FOR PUBLICATION
10 February 2021

PUBLISHED
2 March 2021

LETTER

New approach for predicting nitrification and its fraction of N_2O emissions in global terrestrial ecosystems

Baobao Pan[1], **Shu Kee Lam**[1], **Enli Wang**[2], **Arvin Mosier**[1] **and Deli Chen**[1]

[1] School of Agriculture and Food, Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, VIC 3010, Australia
[2] CSIRO Agriculture and Food, GPO Box 1700, Canberra ACT 2601, Australia

E-mail: shukee.lam@unimelb.edu.au

Keywords: gross nitrification rate, N_2O from nitrification, machine learning, nitrogen cycle, climate change, modelling

Supplementary material for this article is available online

Abstract

Nitrification is a major pathway of N_2O production in aerobic soils. Measurements and model simulations of nitrification and associated N_2O emission are challenging. Here we innovatively integrated data mining and machine learning to predict nitrification rate (R_{nit}) and the fraction of nitrification as N_2O emissions ($f_{N_2O_{Nit}}$). Using our global database on R_{nit} and $f_{N_2O_{Nit}}$, we found that the machine-learning based stochastic gradient boosting (SGB) model outperformed three widely used process-based models in estimating R_{nit} and N_2O emission from nitrification. We then applied the SGB technique for global prediction. The potential R_{nit} was driven by long-term mean annual temperature, soil C/N ratio and soil pH, whereas $f_{N_2O_{Nit}}$ by mean annual precipitation, soil clay content, soil pH, soil total N. The global $f_{N_2O_{Nit}}$ varied by over 200 times (0.006%–1.2%), which challenges the common practice of using a constant value in process-based models. This study provides insights into advancing process-based models for projecting N dynamics and greenhouse gas emissions using a machine learning approach.

1. Introduction

Nitrous oxide (N_2O) is a potent greenhouse gas and can cause ozone depletion (Ravishankara *et al* 2009, Robertson and Vitousek 2009). Nitrification, a microbial process that converts ammonium (NH_4^+) into nitrite and subsequently nitrate, is a major pathway of N_2O production, especially in aerobic soils. However, measurements of nitrification rate (R_{nit}) and how much nitrified N is emitted as N_2O are challenging. Many ecological process-based models, such as APSIM (Keating *et al* 2003, Holzworth *et al* 2014), DNDC (Li *et al* 1992, 2000, Zhang *et al* 2002), WNMM (Li et al 2007), and DAYCENT (Parton *et al* 1998, Del Grosso *et al* 2000) have embedded modules to predict R_{nit} and N_2O from nitrification. These models adopt the equations generated from limited empirical data, and nitrification is calculated as a function of soil water content, soil pH, soil temperature and soil NH_4^+ content. The models adopt a 'grey-box' to estimate N_2O emission from nitrification by assuming the fraction of nitrification as N_2O emissions ($f_{N_2O_{Nit}}$) is fixed, and further constraining it by soil water content or soil temperature. The use of a fixed fraction is known to be problematic in predicting N_2O emission because the proportion of N_2O emission from nitrification varies with soil and environmental conditions (Farquharson 2016). In addition, process-based models may not perform well when extrapolating from site-specific to a larger scale, mainly owing to their deficiency in capturing the key processes in response to driving factors, detailed parameterization, limited data availability and their limited capacity in handling complex interacting factors (Giltrap *et al* 2015, Leng and Hall 2020, Saha *et al* 2021).

Because of the nonlinearity between input and output variables and intricate interactions among these input variables (Butterbach-Bahl *et al* 2013), robust prediction of R_{nit} and associated N_2O production is difficult due to uncertainty in existing process-based models and limited measured data to improve those models. To address this issue, we adopted machine learning [stochastic gradient boosting (SGB) modelling] with data technologies and high-performance computing to identify patterns from

global datasets and predict R_{nit} and $f_{N_2O_{Nit}}$. Models based on machine learning, such as ANN (artificial neural networks), RF (random forest), SVR (support vector regression) and BRT (boosted regression tree) have been recently applied in predicting soil biogeochemical processes. Previous studies showed that machine learning models performed better than traditional models in predicting N_2O and NO emission locally (Villa-Vialaneix *et al* 2012, Philibert *et al* 2013, Saha *et al* 2021) and globally (Delon *et al* 2007, Zhuang *et al* 2012, Perlman *et al* 2014), soil nitrogen (N) loss with biochar application (Liu *et al* 2019), soil organic carbon, and total N (Morellos *et al* 2016, Yang *et al* 2016, Schillaci *et al* 2017). Among the machine-learning based models, BRT (Friedman 2001, 2002, Elith *et al* 2008) relied on SGB, which is especially effective when quantifying nonlinear N transformations that are regulated by variables with complicated interactions, like R_{nit} and associated N_2O emissions. Furthermore, SGB is useful in handling relatively small datasets ($n < 100$) by adjusting the weights of training datasets and tuning learning rate based on the size of the datasets to reduce errors and overfitting (Shepherd *et al* 2003, Andonie 2010, Zhang and Ling 2018). As a variable selection process, SGB also excludes less important variables to achieve better generalization of the dependent variable (Xu *et al* 2014). From the functional perspective, process-based models need to define the function of each input variable. On the contrary, response functions and variable interactions could be handled in SGB models, because they do not pre-set response functions (e.g. linear or exponential) but assume a certain interaction between variables (Leng and Hall 2020). Thus, we hypothesized that these features of SGB models could improve prediction performances of R_{nit} and associated N_2O emissions at a large scale.

In this study, using a comprehensive database compiled from global literature, we first attempted to establish SGB models derived from this database to predict R_{nit} and $f_{N_2O_{Nit}}$, and to compare their performance with existing process-based models. The objectives of the study were: (a) to compare the performance of SGB and process-based models in predicting nitrification rate and associated N_2O production (b) to extend spatially the simulation of R_{nit} and $f_{N_2O_{Nit}}$ using the better performing model.

2. Methods

2.1. Database compilation

Extensive keyword searches of databases (Web of Science (ISI), SCOPUS, CAB Abstracts (ISI), Academic Search Complete (EBSCO) and Google Scholar) and the reference list of cited references were performed. The keywords used in the search were nitrification; N_2O/nitrous oxide emission/pathways; agriculture; cropping; pastures; forest and their combinations. Studies were included if they met the following criteria: (a) ^{15}N tracers or acetylene blockage technique was used to determine N_2O emission produced during nitrification; (b) gross nitrification rates (R_{nit}) were provided; (c) paired nitrification rates and associated N_2O emission were reported; (d) details on experimental location, design and conditions were given to enable cross-checking for duplicate publication; (e) instantaneous measurements were excluded due to high heterogeneity of outcomes. We identified 186 observations from 25 papers published prior to 2018 that satisfied the set criteria for this study. These observations were used to train the SGB models. All studies were conducted under laboratory incubation conditions probably owing to the technical limitations of simultaneous measurements of R_{nit} and $f_{N_2O_{Nit}}$ under field conditions. Data were log transformed to reduce the impact of outliers and improve normality for further analysis.

2.2. Comparing the performance of process-based models and SGB models

The algorithms in APSIM (Holzworth *et al* 2014), Crop-DNDC (Zhang *et al* 2002) and WNMM (Li *et al* 2007) simulate R_{nit} using the same input variables, viz. soil NH_4^+ content, soil pH, soil water content and soil temperature. The responses of nitrification to soil pH, soil water content and soil temperature during nitrification vary between these three models (figure 1). Crop-DNDC covers a wide range of soil water conditions with an increased response up to 80% WFPS. In APSIM and WNMM models, the thresholds of soil water content for nitrification are higher than that in Crop-DNDC. Nitrification increases with soil temperature in WNMM and APSIM models but decreases in Crop-DNDC when soil temperature is over 35 °C. These models estimate N_2O production from nitrification through multiplying the simulated nitrification rate by a fixed $f_{N_2O_{Nit}}$, and/or combining with functions of WFPS and soil temperature (figure 1). Specifically, APSIM simply adopts a simple default $f_{N_2O_{Nit}}$ of 0.002. WNMM considers soil temperature and soil water content for the calculation but sets a threshold $f_{N_2O_{Nit}}$ of 0.002. Crop-DNDC includes soil WFPS and temperature and adopts a constant $f_{N_2O_{Nit}}$ of 0.0006. The compiled database (section 2.1) was used to calculate the nitrification rate and N_2O from nitrification based on the equations (see supplementary information 1 (available online at stacks.iop.org/ERL/16/034053/mmedia)) of WNMM, Crop-DNDC and APSIM.

A SGB model is a tree-based model; each tree is built in a sequential error-correcting process to converge to an accurate model. The SGB model was proposed and modified based on the gradient boosting decision tree (Breiman 1996, Friedman 2001, 2002). SGB model includes random subsampling, which draws a subsample of training data randomly instead

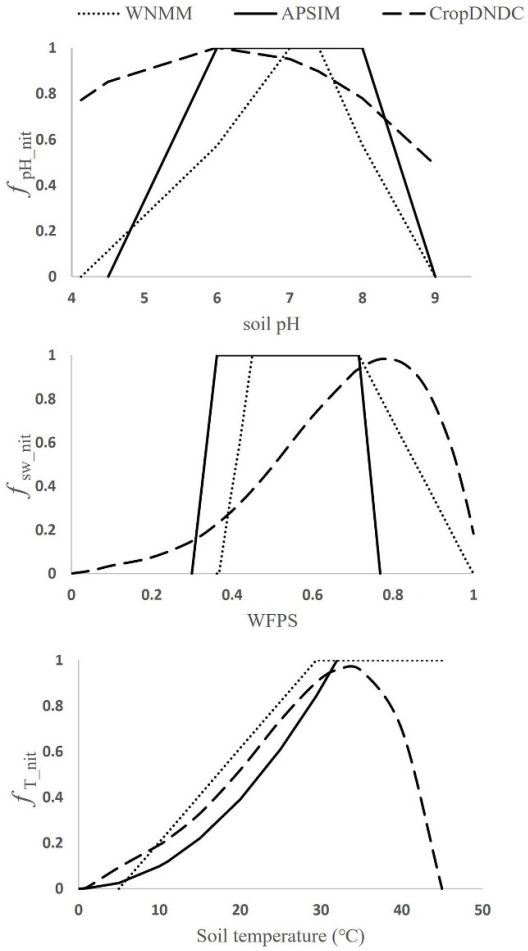

Figure 1. The effects of soil pH, soil temperature and WFPS on nitrification in process-based models (ASPIM, WNMM and CropDNDC).

of boosting all the sample data at each iteration (see supplementary information 3). This modification reduces the computational complexity and improves the model learning speed. The SGB model can also handle both numerical and categorical variables as features (Friedman 2001, 2002, Wang *et al* 2017). To compare with the three process-based models we also adopted the same input variables, i.e. soil NH_4^+ content, soil pH, soil water content and soil temperature to build an SGB model, referred to as SGB1. Four parameters that can be configured in a SGB model are the learn rate (LR), subsample fraction (SF), number of trees (NT), and minimum terminal node (MINCHILD). Learn rate controls the contribution of each tree to the final model. Subsample fraction is involved in the learning process to prevent model from overfitting. The number of trees is based on the learn rate. Minimum terminal nodes are regulated by the size of dataset. To perform the regularization, we tested different combinations of LR (0.001, 0.005, 0.01), SF (0.3, 0.6, 0.8, 0.9), NT (1000, 2000, 3000) and MINCHILD (3, 6, 9) parameters, and found that the optimal configurations for SGB1 were 0.01, 0.8, 3000 and 3 for LC, SF, NT and MINCHILD, respectively (Friedman 2001, Yang *et al* 2016, Wang *et al* 2018) (figure S1). The accuracy of the model was evaluated by 10-fold cross-validation (CV), where the entire dataset was first used for learning purposes, and subsequently partitioned into ten bins. For each of the 10 folds, nine folds were used as a training set and the remaining fold as a test set. The test results from each fold were averaged to estimate the whole model performance. CV is particularly useful for small datasets when one cannot afford to reserve some data for testing (Kohavi 1995) (figure S2).

Regression coefficients of determination (R^2), root mean square error (RMSE) and the Nash–Sutcliffe model efficiency (NSE) (Nash and Sutcliffe 1970) (see supplementary information 2), were used to measure the percentage of variation explained by the model and the model accuracy for SGB1 and the three process-based models.

2.3. Prediction of R_{nit} and $f_{N_2O_{Nit}}$ under different soil properties and climate conditions

We attempted to extend the prediction of R_{nit} and $f_{N_2O_{Nit}}$ to a larger scale by applying SGB technique to our global database. It is unrealistic to perform dynamic prediction of R_{nit} and $f_{N_2O_{Nit}}$ owing to data limitation. Previous studies proved that soil properties (clay content, pH, etc) and climatic conditions (MAT and MAP) can affect microbial community biomass and composition in the soil (Avrahami *et al* 2003, Hu *et al* 2016, Liu *et al* 2017), which are the potential key drivers of the nitrification process and associated N_2O emissions. We therefore performed R_{nit} and $f_{N_2O_{Nit}}$ prediction using long-term edaphic and environmental conditions as input variables.

We followed a series of stepwise procedures (Thompson 1978, Moisen *et al* 2006) to eliminate redundant input variables and optimize the number of variables for this SGB model (referred to as SGB2). This improves the applicability of SGB2 without compromising much prediction capacity. The stepwise variable selection method is efficient in computation and has been widely used in machine learning variable selection (Guyon *et al* 2002, Rakotomamonjy 2003). Since SGB2 was built based on incubation studies conducted under *in vitro* conditions mostly with optimal soil temperature, soil moisture and N availability, the prediction reflects R_{nit} and $f_{N_2O_{Nit}}$ under such conditions.

We further included a set of constraints to regulate the prediction results by SGB2. We first set soil total N and soil organic carbon as trigger conditions, which are the prerequisites for nitrification. Second, desert areas with low precipitation show limited potential for nitrification (Noy-Meir 1974, Skujiņš 1981). In this regard, typical desert areas (mean annual precipitation (MAP) <250 mm) (Marshak 2010) have been excluded from the prediction. To adjust the deviation

between incubation temperature and long-term temperature, we incorporated a function of mean annual temperature (MAT) against nitrification rate to reflect the *in situ* situations globally. The complete model is described as follows:

$$R_{\mathrm{nit}} = M_{\mathrm{SGB2}} \cdot \gamma = \prod_{\alpha \in F} \delta_\alpha \cdot f(T) \cdot M_{\mathrm{SGB2}} \quad (1)$$

$$\gamma = \delta_F \cdot f(T) \quad (2)$$

$$f(T) = \begin{cases} 0 & \text{if } T < 2.8 \\ 0.0023T2 + 0.0032T - 0.006 & \text{if } 2.8 \leqslant T < 8 \\ 0.0006T^{2.2188} & \text{if } 8 \leqslant T < 28 \\ 1 & \text{if } T \geqslant 28 \end{cases} \quad (3)$$

(Malhi and McGill 1982)

$$\delta_F = \prod_{\alpha \in F} \delta_\alpha \quad (4)$$

$$\delta_{\mathrm{SOC}} = \begin{cases} 0 & \text{if SOC} = 0 \\ 1 & \text{if SOC} > 0 \end{cases} \quad (5)$$

$$\delta_{\mathrm{TN}} = \begin{cases} 0 & \text{if TN} = 0 \\ 1 & \text{if TN} > 0 \end{cases} \quad (6)$$

$$\delta_P = \begin{cases} 0 & \text{if } P < 250 \\ 1 & \text{if } P \geqslant 250 \end{cases} \quad (7)$$

Among these equations, M_{SGB} is the original SGB2 prediction, γ is the constraint coefficient of TN, SOC, MAP (P) and MAT (T), $F = \{\mathrm{TN}, \mathrm{SOC}, P\}$ are a set of total N, soil organic carbon and MAP. δ_F is the constraint coefficient of F, δ_α is the constraint coefficient involved in δ_F, where $\delta_\alpha \in \{\delta_{\mathrm{SOC}}, \delta_{\mathrm{TN}}, \delta_P\}$, and $f(T)$ is the function to determine the impact of MAT on R_{nit}. Apart from R^2, RMSE and NSE, additional mean absolute error (MAE) was used to measure the accuracy of prediction for selecting the optimal SGB model when mapping global prediction of R_{nit} and $f_{\mathrm{N_2O_{Nit}}}$.

To project R_{nit} and $f_{\mathrm{N_2O_{Nit}}}$ at a global scale using SGB, potential driving factors of soil and climate data were obtained from different sources and integrated into a spatial GIS database. The soil property database with a spatial resolution of 1 km^2 was obtained from the World Inventory of Soil Emission (WISE) database developed by the International Soil Reference and Information Centre (ISRIC) (www.isric.org) (Batjes 2015). The climate data at 0.5° resolution was collected from Climatic Research Unit (CRU TS v4.01) (www.cru.uea.acuk/data/) from 1911 to 2010 (Harris *et al* 2014). The ESRI ArcGIS software (v. 10.4.1 ESRI) was used to plot the global distribution map. In this study, all SGB models were constructed using TreeNet® (Salford Systems).

3. Results

3.1. Overview of the literature-based dataset

Study sites spanned from 122° W to 152° E and 43° S to 65° N. For most of the incubation experiments, the soil temperature was controlled at 20 °C–25 °C, with a range of 5 °C–45 °C, and average soil moisture of around 45% WFPS. The average R_{nit} in the topsoil (0–20 cm) was 1.4 kg N ha^{-1} d^{-1} and varied widely. The average $f_{\mathrm{N_2O_{Nit}}}$ was 0.46% (0.004%–9.19%) (table 1 and figure S3).

3.2. Comparison of model performance

SGB1 outperformed the other three process-based models with the highest R^2 and NSE value and lowest RMSE in predicting nitrification (0.73, 0.63 and 0.4, respectively) and N_2O emission during nitrification (0.61, 0.54 and 0.61, respectively) (table 2). The R^2 values of R_{nit} for APSIM, Crop-DNDC and WNMM were 0.27, 0.31 and 0.24, respectively. WNMM and Crop-DNDC overestimated nitrification rate by 1.45 kg N ha^{-1} d^{-1} and 3.39 kg N ha^{-1} d^{-1}, respectively, whereas APSIM underestimated it by 1.72 kg N ha^{-1} d^{-1} (figure S4). As for the N_2O emission during nitrification calculated using a fixed ratio and functions of soil water content and soil temperature, all three process-based models did not perform very well (table 2). Crop-DNDC overestimated N_2O emission from nitrification by 1.08 kg N ha^{-1} d^{-1}. APSIM and WNMM underestimated N_2O emission when it was <0.3 g N ha^{-1} d^{-1} but overestimated it when the N_2O emission from nitrification was higher (figure S5).

3.3. Stepwise variable selection of optimal models

Nitrification rate showed a strong relationship with MAT and soil C/N ratio in the model A1 (table 3). The highest R^2 value of 0.8 was achieved when at least six variables were included (models A5–A8). When only MAT, soil C/N ratio and soil pH were included (model A2), the performance reached 0.76 for both R^2 and NSE (table 3), which we considered optimal for further upscale predictions when compromising between the R^2 value and the number of variables included.

When predicting $f_{\mathrm{N_2O_{Nit}}}$, the performance of model B3 was significantly increased with four input variables (MAP, soil clay content, soil pH and soil TN), yielding a relatively higher R^2 value (0.55) and NSE (0.55), and lower MAE (0.26) and RMSE (0.40) than other models (table 4). Therefore, B3 model (table 2, SGB2) was considered optimal for predicting $f_{\mathrm{N_2O_{Nit}}}$.

3.4. Global mapping of the response of R_{nit} and $f_{\mathrm{N_2O_{Nit}}}$ to soil properties and climate conditions

We extend the prediction of R_{nit} spatially using SGB2 (model A2 in table 3) with input variables soil pH,

Table 1. Summary statistics of nitrification, the fraction of associated N_2O emissions and environmental variables from global literature.

	n	Mean	Median	Min	Max	SD
R_{nit} (kg N ha^{-1} d^{-1})	186	1.40	0.29	0.003	27.5	3.9
$f_{N_2O_{Nit}}$ (%)	186	0.46	0.12	0.004	9.20	1.18
MAP (mm)	186	887	700	423	2416	470
MAT (°C)	186	13.5	14.3	−0.6	25.0	5.0
Soil temperature (°C)	181	21.3	22.0	5.0	45.0	5.2
Soil clay content (%)	139	23.2	19.4	4.0	60.6	16.4
Soil pH	186	6.4	6.4	3.7	8.4	1.2
SOC (%)	185	2.4	2.0	1.0	9.5	1.7
Soil TN (%)	114	0.28	0.18	0.09	2.23	0.34
Soil C/N ratio	116	11.7	11.5	2.3	22.2	3.8
Soil NH_4^+–N (mg kg^{-1})	154	18.1	13.0	0.3	116.0	22.1
Soil NO_3^-N (mg kg^{-1})	154	35.9	11.0	1.0	1157.0	120.7
Soil WFPS (%)	174	44.5	42.0	24.0	90.0	13.1
N input (kg N ha^{-1})	186	23.7	8.7	0.0	300.0	48.3

soil C/N ratio and MAT from global soil and climate databases. Nitrification rate varied spatially, spanning a wide range of 0.0001–2.19 kg N ha^{-1} d^{-1} with an average of 0.11 kg N ha^{-1} d^{-1}(figure 2). The highest R_{nit} was observed in subtropical and equatorial areas. Higher nitrification rates generally occur in soils with a lower C/N ratio and a neutral pH (figure S6). $f_{N_2O_{Nit}}$ covered a broad range from 0.006% to 1.24% (figure 3) with an average of 0.13%. Lower $f_{N_2O_{Nit}}$ was noted in tropical and higher latitude regions in the northern hemisphere, with higher $f_{N_2O_{Nit}}$ across the subtropical and around equatorial areas. A large proportion of N_2O from nitrification was found in soils with a lower clay content (figure S7). When soil is neutral pH, more N_2O was released from nitrification compared to both acidic and alkaline soils (figure S7). $f_{N_2O_{Nit}}$ followed the pattern of MAP and soil total N (figure S7).

4. Discussion

4.1. Better performance of SGB models

SGB1 better predicts nitrification and associated N_2O emissions than the three widely used process-based models using the same input variables (soil NH_4^+ content, soil pH, soil water content and soil temperature). There are several possible reasons. First, limited site-specific experimental data were used for deriving process-based model parameters and equations, where less responsive or redundant variables might have been included. On the contrary, the relationships between variables in SGB1 were more representative as their internal interactions were developed from a relatively comprehensive global database.

Second, process-based models may have misused the variable responses for R_{nit} and $f_{N_2O_{Nit}}$ in their equations. For example, the relatively narrow range of the response of soil water to nitrification limited the prediction capacity of the APSIM and WNMM models. If soil water content was beyond of the lower and upper limits set in the equations in these models, nitrification was assumed to stop. In particular we found that APSIM and WNMM estimated that no nitrification occurred below 30% WFPS. However, the measured data from our global database indicated that nitrification did occur under 30% WFPS (Maag and Vinther 1996).

Third, the equations currently used in process-based models cannot account for the interactions of input variables (Butterbach-Bahl *et al* 2013) without comprehensive calibration using more site-specific detailed data. In contrast, SGB1 examines all possible nonlinear relationships and interactions between input variables themselves and output variables (Ryo and Rillig 2017).

4.2. The use of SGB models in spatial prediction globally

The spatial patterns of predicted R_{nit} and $f_{N_2O_{Nit}}$ exhibited a large variation owing to the diverse soil and environmental conditions and their interactions (figures 2 and 3). This result further shows that using a constant $f_{N_2O_{Nit}}$ to estimate the N_2O emission from nitrification by existing process-based models is unsuitable (Chen *et al* 2008, Farquharson 2016).

The relationship between soil properties and R_{nit} has been well studied (Dancer *et al* 1973, Nyborg *et al* 1988, Bengtsson *et al* 2003, Zebarth *et al* 2015). A recent global-scale study reported that soil C/N ratio, soil pH and MAT are the key drivers of R_{nit} (Li *et al* 2020). High soil C/N ratio may stimulate N immobilization, decrease availability of NH_4^+ for nitrification, and subsequently lower nitrification rate (Bengtsson *et al* 2003). Larger microbial populations and activities (including nitrifiers) are generally found in neutral rather than in high or low pH soils (Tabatabai *et al* 1992). MAT can regulate soil nitrification rate directly and indirectly by changing microbial community structure and abundance in the long term (Avrahami *et al* 2003, Hu *et al* 2016). Besides, MAT

Table 2. Comparison of observed and predicted nitrification and associated N_2O emissions by SGB1, APSIM, Crop-DNDC and WNMM models.

	SGB1				APSIM				Crop-DNDC				WNMM			
Output	*n*	R^2	RMSE	NSE	*n*	R^2	RMSE	NSE	*n*	R^2	RMSE	NSE	*n*	R^2	RMSE	NSE
Log_Nit[a]	186	0.73	0.40	0.63	99	0.27	0.71	−0.02	162	0.31	0.68	−1.04	98	0.24	0.66	−0.45
Log_N_2O_Nit[b]	186	0.61	0.54	0.61	99	0.10	0.84	−0.51	162	0.20	0.73	−14.2	98	0.11	0.68	−0.89

[a] Nitrification rate.

[b] N_2O emission from nitrification, RMSE: root mean square error and NSE: the Nash–Sutcliffe model efficiency.

Table 3. Model prediction of R_{nit} by stepwise selection of variables.

Model	Stepwise variable selection	R^2	MAE	RMSE	NSE
A1	MAT, soil C/N ratio	0.73	0.29	0.40	0.73
A2	MAT, soil C/N ratio, soil pH	0.76	0.26	0.37	0.76
A3	MAT, MAP, soil C/N ratio, soil pH	0.76	0.26	0.37	0.76
A4	Land use, MAT, MAP, soil C/N ratio, soil pH	0.79	0.25	0.35	0.79
A5	Land use, MAT, MAP, soil C/N ratio, soil pH, SOC	0.80	0.25	0.34	0.80
A6	Land use, MAT, MAP, soil C/N ratio, soil pH, SOC, soil TN	0.80	0.25	0.34	0.80
A7	Land use, MAT, MAP, soil C/N ratio, soil pH, SOC, soil TN, soil clay content	0.80	0.25	0.34	0.80

Table 4. Model prediction of $f_{N_2O_{Nit}}$ by stepwise selection of variables.

Model	Stepwise variable selection	R^2	MAE	RMSE	NSE
B1	MAP, soil clay content	0.39	0.30	0.48	0.35
B2	MAP, soil clay content, soil pH	0.44	0.30	0.46	0.41
B3	MAP, soil clay content, soil pH, soil TN	0.55	0.26	0.40	0.55
B4	MAP, soil clay content, soil pH, soil TN, soil C/N ratio	0.57	0.26	0.39	0.57
B5	Land use, MAP, soil clay content, soil pH, soil TN, soil C/N ratio	0.58	0.26	0.39	0.58
B6	Land use, MAP, MAT, soil clay content, soil pH, soil TN, soil C/N ratio	0.57	0.26	0.39	0.56
B7	Land use, MAP, MAT, soil clay content, soil pH, SOC, soil TN, soil C/N ratio	0.54	0.27	0.41	0.53

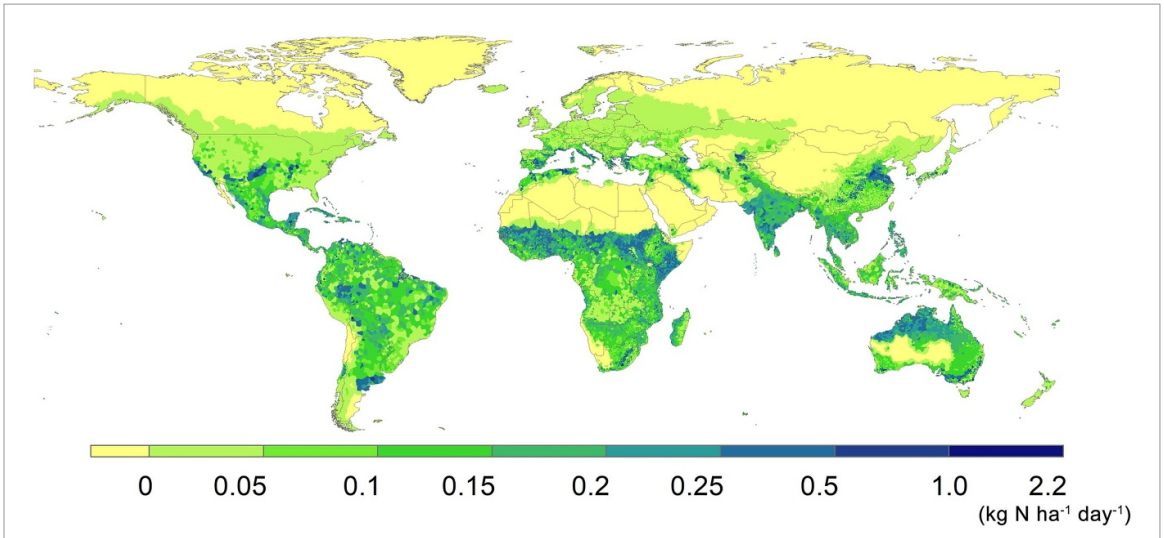

Figure 2. Global distribution of R_{nit} predicted by SGB (A2) model.

reflects the environmental conditions on site and regulates the long-term soil microclimate.

The spatial distribution of $f_{N_2O_{Nit}}$ was consistent with land use distribution (figure S8), which reflects the edaphic and climatic variations across terrestrial systems. The estimated range of $f_{N_2O_{Nit}}$ in humid tropical areas by B3 (table 4) was comparable to the DNDC and DLEM models examined by Inatomi *et al* (2019), but much lower than VISIT (Inatomi *et al* 2010) model estimates. However, $f_{N_2O_{Nit}}$ in semiarid regions was predicted to be lower by DNDC, DELM and VISIT than the B3 model. The differences in model predictions can be explained by three reasons. First, Inatomi *et al* (2019) adopted common protocols and initial and boundary conditions when upscaling the simulation for all three process-based models. Second, for the VISIT model, the parameterization of $f_{N_2O_{Nit}}$ was only dependent on soil pH, which is insufficient to reflect the complicated relationship between $f_{N_2O_{Nit}}$ and soil and environmental factors. Third, soil moisture and temperature in semiarid regions were below the threshold set for nitrification to occur

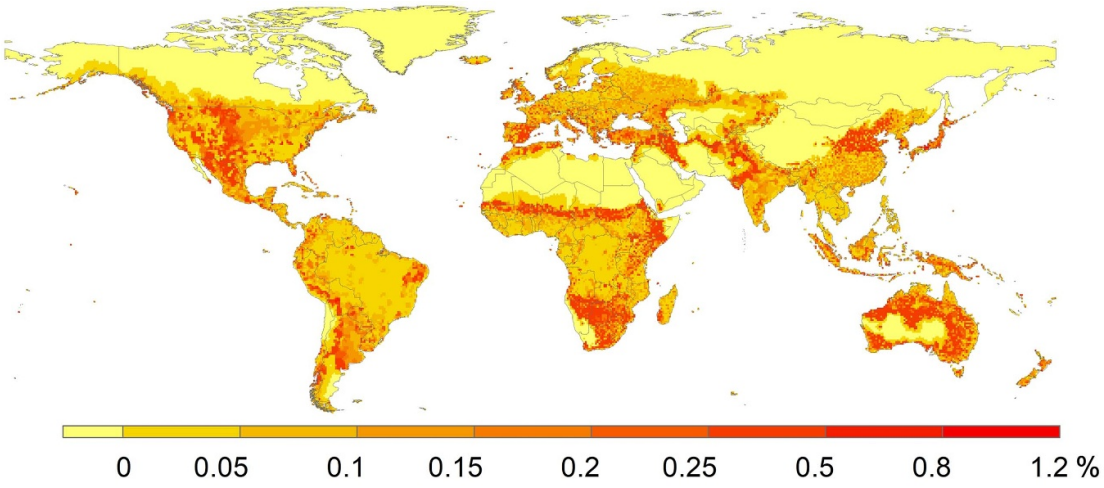

Figure 3. Global distribution of $f_{N_2O_{Nit}}$ predicted by SGB (B3) model.

in process-based models, thereby underestimating nitrification and associated N_2O emissions, whereas our A2 and B3 models cover a wider range of input variables.

In our study, subtropical areas and equatorial areas have higher R_{nit} and $f_{N_2O_{Nit}}$ than other areas. Higher MAT, neutral soil pH and lower soil C/N ratio could increase the microbial populations in soils and nitrifier efficiency, thus increasing nitrification rate (Gilmour 1984, Tabatabai *et al* 1992). The highest R_{nit} and $f_{N_2O_{Nit}}$ were observed in areas with intense human activities, i.e. cropland and grasslands (Ambus 1998). These could be attributed to long-term anthropogenic N input, increasing soil N availability for nitrification and associated N_2O production (Tabatabai *et al* 1992). On the other hand, optimal MAP provided an ideal oxygen level and moderate moisture condition for nitrifiers (Yu and Zhuang 2019). Areas with extensive forest and savannas have the lowest R_{nit} and $f_{N_2O_{Nit}}$ worldwide. This finding indicates that in temperate and tropical forest areas, nitrification is not a major source of N_2O emissions, which is in agreement with previous studies (Stehfest and Bouwman 2006, Werner *et al* 2007, Cheng *et al* 2012, Zhuang *et al* 2012).

Responses and changes of R_{nit} and $f_{N_2O_{Nit}}$ are highly related to long-term environmental, edaphic factors and land use. The fraction of nitrification as N_2O emissions is clearly not a constant; instead it should be adjusted according to edaphic and environmental conditions when used in process-based models or global climate models in projecting N_2O emissions. We are aware of the limited datasets used to develop the SGB2 models for global prediction and did not intend to accurately quantify R_{nit} and $f_{N_2O_{Nit}}$ at any timepoint. Our objective was to demonstrate that SGB2 models can be used to map the global spatial patterns of potential R_{nit} and $f_{N_2O_{Nit}}$ based on a few long-term soil and climate variables that are easily accessible from world databases. These potential R_{nit} and $f_{N_2O_{Nit}}$ values obtained under optimal soil temperature, soil moisture and sufficient N availability, could be used as benchmark for the calibration of process-based models.

Our findings provide important implications for the prediction of the N losses in process-based models. The SGB model is able to derive new parameters that can be incorporated into process-based models for different N loss pathways under various edaphic and environmental conditions. Moreover, SGB could also be directly embedded in a process-based model or replace the existing unsatisfactory modules of a process-based model to capture the complex, dynamic, and nonlinear processes more efficiently and effectively. By integrating with other modules of process-based models, SGB models can be used to develop decision support tools for sustainable N management.

4.3. Limitation

Limitations remained in this study. First, we conducted a comprehensive global literature search, but found a low number of observations. This indicates limited geographical coverage and the difficulties of simultaneously measuring nitrification and associated N_2O emissions both *in situ* and in the laboratory. Second, while SGB models are suitable for handling small databases, the inclusion of more data in training machine learning models would likely improve their performance. Third, although we included constraint coefficients in SGB models to address the issues associated with artificial experimental conditions, this potential data bias can be reduced when more *in situ* data become available.

5. Conclusion

This study is the first attempt to use machine learning to develop SGB models to predict R_{nit} and $f_{N_2O_{Nit}}$ in terrestrial ecosystems. Compared to three widely used process-based models, the SGB models were more accurate in predicting nitrification rate and associated N_2O emissions. The SGB models can be extended to a global level with only a few input variables without compromising accuracy. In particular, nitrification rate was predicted by soil pH, MAT and soil C/N ratio whereas the fraction of nitrification as N_2O emissions was predicted using soil pH, MAP, soil total N and clay content. Large spatial variation in nitrification and its fraction as N_2O emissions was mainly driven by long-term environmental and edaphic factors. The fraction of nitrification as N_2O emissions is clearly not constant; instead it should be adjusted according to edaphic and environmental conditions when used in process-based models or global climate models in projecting N_2O emissions.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The authors acknowledge support from Australian Research Council Linkage Project (LP160101417), Australian Government Research Training Program Scholarship, Leslie H Brunning Research Scholarship and the Australia-China Joint Research Centre—Healthy Soils for Sustainable Food Production and Environmental Quality (ACSRF48165), and advice on the machine-learning technique from Usha Nattala and Geordie Zhang of the Melbourne Data Analytics Platform of the University of Melbourne.

ORCID iDs

Baobao Pan https://orcid.org/0000-0003-0147-8471
Shu Kee Lam https://orcid.org/0000-0001-7943-5004
Deli Chen https://orcid.org/0000-0001-6767-1376

References

Ambus P 1998 Nitrous oxide production by denitrification and nitrification in temperate forest, grassland and agricultural soils *Eur. J. Soil Sci.* **49** 495–502
Andonie R 2010 Extreme data mining: inference from small datasets *Int. J. Comput. Commun. Control* **5** 280–91
Avrahami S, Liesack W and Conrad R 2003 Effects of temperature and fertilizer on activity and community structure of soil ammonia oxidizers *Environ. Microbiol.* **5** 691–705
Batjes N H 2015 *World Soil Property Estimates for Broad-scale Modelling WISE30sec* (ISRIC-World Soil Information)
Bengtsson G, Bengtson P and Mansson K F 2003 Gross nitrogen mineralization, immobilization, and nitrification rates as a function of soil C/N ratio and microbial activity *Soil Biol. Biochem.* **35** 143–54
Breiman L 1996 Bagging predictors *Mach. Learn.* **24** 123–40
Butterbach-Bahl K, Baggs E M, Dannenmann M, Kiese R and Zechmeister-Boltenstern S 2013 Nitrous oxide emissions from soils: how well do we understand the processes and their controls? *Phil. Trans. R. Soc.* B **368** 20130122
Chen D, Li Y, Grace P and Mosier A R 2008 N_2O emissions from agricultural lands: a synthesis of simulation approaches *Plant Soil* **309** 169–89
Cheng Y, Cai Z C, Zhang J B, Lang M, Mary B and Chang S X 2012 Soil moisture effects on gross nitrification differ between adjacent grassland and forested soils in central Alberta, Canada *Plant Soil* **352** 289–301
Dancer W, Peterson L and Chesters G 1973 Ammonification and nitrification of N as influenced by soil pH and previous N treatments *Soil Sci. Soc. Am. J.* **37** 67–69
Del Grosso S J, Parton W J, Mosier A R, Ojima D S, Kulmala A E and Phongpan S 2000 General model for N_2O and N_2 gas emissions from soils due to dentrification *Glob. Biogeochem. Cycles* **14** 1045–60
Delon C, Serça D, Boissard C, Dupont R, Dutot A, Laville P, De Rosnay P and Delmas R 2007 Soil NO emissions modelling using artificial neural network *Tellus* B **59** 502–13
Elith J, Leathwick J R and Hastie T 2008 A working guide to boosted regression trees *J. Anim. Ecol.* **77** 802–13
Farquharson R 2016 Nitrification rates and associated nitrous oxide emissions from agricultural soils—a synopsis *Soil Res.* **54** 469–80
Friedman J H 2001 Greedy function approximation: a gradient boosting machine *Ann. Stat.* **29** 1189–232
Friedman J H 2002 Stochastic gradient boosting *Comput. Stat. Data Anal.* **38** 367–78
Gilmour J 1984 The effects of soil properties on nitrification and nitrification inhibition *Soil Sci. Soc. Am. J.* **48** 1262–6
Giltrap D, Vogeler I, Cichota R, Luo J, Van Der Weerden T and De Klein C 2015 Comparison between APSIM and NZ-DNDC models when describing N-dynamics under urine patches *N.Z. J. Agric. Res.* **58** 131–55
Guyon I, Weston J, Barnhill S and Vapnik V 2002 Gene selection for cancer classification using support vector machines *Mach. Learn.* **46** 389–422
Harris I, Jones P D, Osborn T J and Lister D H 2014 Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 dataset *Int. J. Climatol.* **34** 623–42
Holzworth D P, Huth N I, Zurcher E J, Herrmann N I, McLean G, Chenu K, Van Oosterom E J, Snow V, Murphy C and Moore A D 2014 APSIM—evolution towards a new generation of agricultural systems simulation *Environ. Model. Softw.* **62** 327–50
Hu H, Macdonald C A, Trivedi P, Anderson I C, Zheng Y, Holmes B, Bodrossy L, Wang J, He J and Singh B K 2016 Effects of climate warming and elevated CO_2 on autotrophic nitrification and nitrifiers in dryland ecosystems *Soil Biol. Biochem.* **92** 1–15
Inatomi M, Hajima T and Ito A 2019 Fraction of nitrous oxide production in nitrification and its effect on total soil emission: a meta-analysis and global-scale sensitivity analysis using a process-based model *PLoS One* **14** e0219159
Inatomi M, Ito A, Ishijima K and Murayama S 2010 Greenhouse gas budget of a cool-temperate deciduous broad-leaved forest in Japan estimated using a process-based model *Ecosystems* **13** 472–83
Keating B A, Carberry P S, Hammer G L, Probert M E, Robertson M J, Holzworth D, Huth N I, Hargreaves J N, Meinke H and Hochman Z 2003 An overview of APSIM, a model designed for farming systems simulation *Eur. J. Agron.* **18** 267–88
Kohavi R 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection *Int. Joint Conf. on*

Artificial Intelligence (IJCAI) (Montreal: IJCAI) pp 1137–45

Leng G and Hall J W 2020 Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models *Environ. Res. Lett.* **15** 044027

Li C, Aber J, Stange F, Butterbach-Bahl K and Papen H 2000 A process-oriented model of N_2O and NO emissions from forest soils: 1. Model development *J. Geophys. Res.: Atmos.* **105** 4369–84

Li C, Frolking S and Frolking T A 1992 A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity *J. Geophys. Res.: Atmos.* **97** 9759–76

Li Y, White R, Chen D, Zhang J, Li B, Zhang Y, Huang Y and Edis R 2007 A spatially referenced water and nitrogen management model WNMM for irrigated intensive cropping systems in the North China Plain *Ecol. Modelling* **203** 395–423

Li Z, Zeng Z, Tian D, Wang J, Fu Z, Zhang F, Zhang R, Chen W, Luo Y and Niu S 2020 Global patterns and controlling factors of soil nitrification rate *Glob. Change Biol.* **26** 1–11

Liu Q, Liu B, Zhang Y, Hu T, Lin Z, Liu G, Wang X, Ma J, Wang H and Jin H 2019 Biochar application as a tool to decrease soil nitrogen losses NH_3 volatilization, N_2O emissions, and N leaching from croplands: options and mitigation strength in a global perspective *Glob. Change Biol.* **25** 2077–93

Liu Y, Wang C, He N, Wen X, Gao Y, Li S, Niu S, Butterbach-Bahl K, Luo Y and Yu G 2017 A global synthesis of the rate and temperature sensitivity of soil nitrogen mineralization: latitudinal patterns and mechanisms *Glob. Change Biol.* **23** 455–64

Maag M and Vinther F P 1996 Nitrous oxide emission by nitrification and denitrification in different soil types and at different soil moisture contents and temperatures *Appl. Soil Ecol.* **4** 5–14

Malhi S and McGill W 1982 Nitrification in three Alberta soils: effect of temperature, moisture and substrate concentration *Soil Biol. Biochem.* **14** 393–9

Marshak S 2010 *Essentials of Geology* 3rd edn (New York: W W.Norton & Company Limited) p 648

Moisen G G, Freeman E A, Blackard J A, Frescino T S, Zimmermann N E and Edwards J T C 2006 Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods *Ecol. Modelling* **199** 176–87

Morellos A, Pantazi X E, Moshou D, Alexandridis T, Whetton R, Tziotzios G, Wiebensohn J, Bill R and Mouazen A M 2016 Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy *Biosyst. Eng.* **152** 104–16

Nash J E and Sutcliffe J V 1970 River flow forecasting through conceptual models part 1—a discussion of principles *J. Hydrol.* **10** 282–90

Noy-Meir I 1974 Desert ecosystems: higher trophic levels *Annu. Rev. Ecol. Evol. Syst.* **5** 195–214

Nyborg M, Hoyt P B and Penney D C 1988 Ammonification and nitrification of N in acid soils at 26 field sites one year after liming *Commun. Soil Sci. Plant Anal.* **19** 1371–9

Parton W J, Hartman M, Ojima D and Schimel D 1998 DAYCENT and its land surface submodel: description and testing *Glob. Planet. Change* **19** 35–48

Perlman J, Hijmans R J and Horwath W R 2014 A metamodelling approach to estimate global N_2O emissions from agricultural soils *Glob. Ecol. Biogeogr.* **23** 912–24

Philibert A, Loyce C and Makowski D 2013 Prediction of N_2O emission from local information with Random Forest *Environ. Pollut.* **177** 156–63

Rakotomamonjy A 2003 Variable selection using SVM-based criteria *J. Mach. Learn. Res.* **3** 1357–70

Ravishankara A R, Daniel J S and Portmann R W 2009 Nitrous oxide N_2O : the dominant ozone-depleting substance emitted in the 21st century *Science* **326** 123–5

Robertson G P and Vitousek P M 2009 Nitrogen in agriculture: balancing the cost of an essential resource *Annu. Rev. Environ. Resour.* **34** 97–125

Ryo M and Rillig M C 2017 Statistically reinforced machine learning for nonlinear patterns and variable interactions *Ecosphere* **8** e01976

Saha D, Basso B and Robertson G P 2021 Machine learning improves predictions of agricultural nitrous oxide (N_2O) emissions from intensively managed cropping systems *Environ. Res. Lett.* **16** 024004

Schillaci C, Lombardo L, Saia S, Fantappiè M, Märker M and Acutis M 2017 Modelling the topsoil carbon stock of agricultural lands with the stochastic gradient treeboost in a semi-arid Mediterranean region *Geoderma* **286** 35–45

Shepherd K D, Palm C A, Gachengo C N and Vanlauwe B 2003 Rapid characterization of organic resource quality for soil and livestock management in tropical agroecosystems using near-infrared spectroscopy *Agron. J.* **95** 1314–22

Skujiņš J 1981 Nitrogen cycling in arid ecosystems *Ecol. Bull.* 477–91 (www.jstor.org/stable/45128683)

Stehfest E and Bouwman L 2006 N_2O and NO emission from agricultural fields and soils under natural vegetation: summarizing available measurement data and modeling of global annual emissions *Nutr. Cycling Agroecosyst.* **74** 207–28

Tabatabai M, Fu M and Basta N 1992 Effect of cropping systems on nitrification in soils *Commun. Soil Sci. Plant Anal.* **23** 1885–91

Thompson M L 1978 Selection of variables in multiple regression: part I. A review and evaluation *Int. Stat. Rev.* **46** 1

Villa-Vialaneix N, Follador M, Ratto M and Leip A 2012 A comparison of eight metamodeling techniques for the simulation of N_2O fluxes and N leaching from corn crops *Environ. Model. Softw.* **34** 51–66

Wang S, Adhikari K, Wang Q, Jin X and Li H 2018 Role of environmental variables in the spatial distribution of soil carbon C, nitrogen N, and C:N ratio from the northeastern coastal agroecosystems in China *Ecol. Indic.* **84** 263–72

Wang S, Zhuang Q, Wang Q, Jin X and Han C 2017 Mapping stocks of soil organic carbon and soil total nitrogen in Liaoning Province of China *Geoderma* **305** 250–63

Werner C, Butterbach-Bahl K, Haas E, Hickler T and Kiese R 2007 A global inventory of N_2O emissions from tropical rainforest soils using a detailed biogeochemical model *Glob. Biogeochem. Cycles* **21** 3

Xu Z, Huang G, Weinberger K Q and Zheng A X 2014 Gradient boosted feature selection *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining 2014* (ACM Press) 522–31

Yang R, Zhang G, Liu F, Lu Y, Yang F, Yang F, Yang M, Zhao Y and Li D 2016 Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem *Ecol. Indic.* **60** 870–8

Yu T and Zhuang Q 2019 Quantifying global N_2O emissions from natural ecosystem soils using trait-based biogeochemistry models *Biogeosciences* **16** 207–22

Zebarth B J, Forge T A, Goyer C and Brin L D 2015 Effect of soil acidification on nitrification in soil *Can. J. Soil Sci.* **95** 359–63

Zhang Y, Li C, Zhou X and Moore B III 2002 A simulation model linking crop growth and soil biogeochemistry for sustainable agriculture *Ecol. Modelling* **151** 75–108

Zhang Y and Ling C 2018 A strategy to apply machine learning to small datasets in materials science *npj Comput. Mater.* **4** 25

Zhuang Q, Lu Y and Chen M 2012 An inventory of global N_2O emissions from the soils of natural terrestrial ecosystems *Atmos. Environ.* **47** 66–75