scientific data

OPEN

DATA DESCRIPTOR

A global dataset of gross nitrogen transformation rates across terrestrial ecosystems

Eunji Byun[1], Christoph Müller[2,3,4], Barbara Parisse[5], Rosario Napoli[5], Jin-Bo Zhang[4,6], Fereidoun Rezanezhad[7], Philippe Van Cappellen[7], Gerald Moser[2,4], Anne B. Jansen-Willems[3,4], Wendy H. Yang[8], Rieko Urakawa[9], José Ignacio Arroyo[10,11], Ulderico Neri[5], Ahmed S. Elrys[4,12,13] & Pierfrancesco Nardi[4,5] ✉

Rates of nitrogen transformations support quantitative descriptions and predictive understanding of the complex nitrogen cycle, but measuring these rates is expensive and not readily available to researchers. Here, we compiled a dataset of gross nitrogen transformation rates (GNTR) of mineralization, nitrification, ammonium immobilization, nitrate immobilization, and dissimilatory nitrate reduction to ammonium in terrestrial ecosystems. Data were extracted from 331 studies published from 1984–2022, covering 581 sites. Globally, 1552 observations were appended with standardized soil, vegetation, and climate data (49 variables in total) potentially contributing to the observed variations of GNTR. We used machine learning-based data imputation to fill in partially missing GNTR, which improved statistical relationships between theoretically correlated processes. The dataset is currently the most comprehensive overview of terrestrial ecosystem GNTR and serves as a global synthesis of the extent and variability of GNTR across a wide range of environmental conditions. Future research can utilize the dataset to identify measurement gaps with respect to climate, soil, and ecosystem types, delineate GNTR for certain ecoregions, and help validate process-based models.

Background & Summary

The soil nitrogen (N) cycle includes several interconnected microbially mediated processes through which N is continuously transformed from one form to another form (Fig. 1). The balance between these processes regulates the availability of N in soil, therefore supporting plant growth, controlling N losses, and ecosystem functioning[1].

Soil N transformations can be measured in terms of net and gross rates. Net rates characterise the overall pool size change as the sum of competing processes of a particular N species. For instance, net mineralization rate quantifies the balance between productive and consumptive ammonium (NH_4^+) processes. On the other hand, measurements of gross rates provide us with unique process specific N rates as their determination allow the quantification of the unidirectional flux between two pools[2]. Gross N process rates are determined using ^{15}N isotope techniques such as the isotope dilution technique[3] and, more recently, ^{15}N tracing techniques based on dilution-enrichment principle[4]. Importantly, gross rates can be several orders of magnitude higher than net rates[5]. Compared to net rate measurements, the determination of gross rates is more expensive, requires

[1]Department of Earth System Sciences, Yonsei University, Seoul, Republic of Korea. [2]Institute for Plant Ecology, Justus Liebig University, Giessen, Germany. [3]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland. [4]Liebig Centre for Agroecology and Climate Impact Research, Giessen, Germany. [5]Council for Agricultural Research and Economics, Research Centre for Agriculture and Environment (CREA-AA), Rome, Italy. [6]School of Geography, Nanjing Normal University, Nanjing, China. [7]Ecohydrology Research Group, Water Institute and Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, ON, Canada. [8]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [9]Asia Center for Air Pollution Research (ACAP), Japan Environmental Sanitation Center (JESC), Niigata, Japan. [10]The Santa Fe Institute, Santa Fe, NM, USA. [11]Center for Mathematical Modeling, Universidad de Chile, Santiago, Chile. [12]School of Tropical Agriculture and Forestry, Hainan University, Haikou, China. [13]Soil Science Department, Faculty of Agriculture, Zagazig University, Zagazig, Egypt. ✉e-mail: pierfrancesco.nardi@crea.gov.it

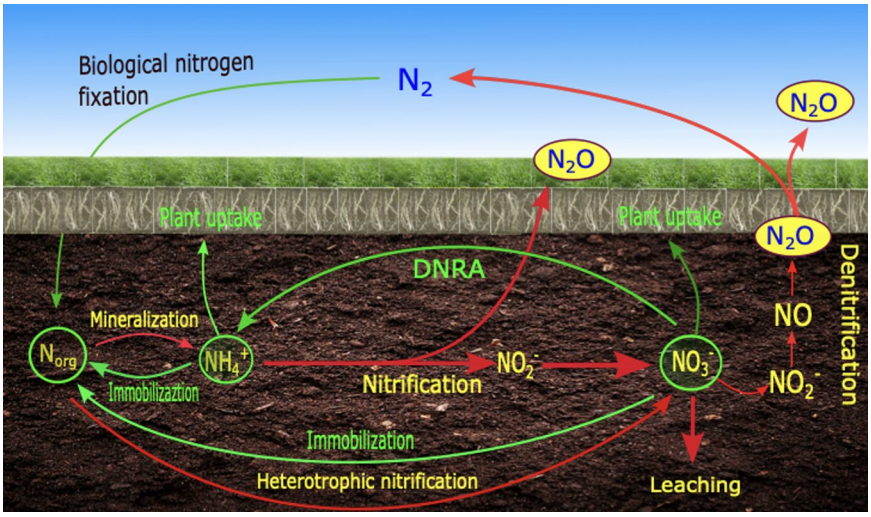

Fig. 1 Main processes of the N cycle. Biological N fixation is described here as the major pathway of N supply to soil environments, referring to the biological uptake and reduction of dinitrogen gas from the atmosphere to bioavailable ammonia (NH_3, not shown) that is highly soluble in soil water as ammonium cation (NH_4^+). Mineralization is another source of soil NH_4^+ increase as it returs organically bound nitrogen (N_{org}) by microbial degradation, also known as ammonification. Nitrification refers to the microbial production of nitrate (NO_3^-) with intermediates carried out by autotrophic and heterotrophic microbes. NH_4^+ and N_{org} are the substrates for the autotrophic and heterotrophic pathways, respectively. During the nitrification process, NH_4^+ is mainly oxidized to nitrite (NO_2^-) and subsequently to nitrate (NO_3^-), but also to a lesser extent to nitrous oxide (N_2O), in which the autotrophic pathways are known to play an important role. As these processes are simultaneously occurring, NH_4^+ and NO_3^- are made available for plant root uptake and microbial biomass assimilation (to N_{org}), referred to here as immobilization (of $NH4^+$ and of NO_3^-, respectively). Dissimilatory NO_3^- reduction to NH_4^+ (DNRA) occurs in low oxygen availablity where NO_3^- is used as the electron acceptor for anaerobic microbial decomposition of organic matter (thus, the microbial reduction of NO_3^- via NO_3^- to NH_4^+), generally considered to maintain bioavailable N in the soil system. By contrast, in denitrification NO_3^- is microbially reduced to gaseous forms (N_2O, NO, and N_2 depending on oxygen availability), which can escape to the atmosphere, hence, causing N loss from the soil.

advanced laboratory equipment and specific technical skills. This is why in the scientific literature net rates predominate over gross rates. Still, quality data of gross N transformation rates are crucial to gain insigths into actual fluxes between N pools and to provide mechanistic insights on environmental factors influencing soil N cycling processes[6–9]. Therefore, the utilization of currently available data will help researchers in gaining fundamental understanding of the soil N cycle.

Here, building upon previous work on global scale synthesis of gross N transformation rates[6–8,10], our global dataset has been updated by adding some global standardized environmental variables and by georeferencing every observation for its accurate soil sampling (or field measurement) location. Thus, enabling future studies to generate a summary of gross N rates for the target N transformation process for the region of interest, and classify them for a specific climate regime or by ecosystem type. For example, there are two sets of climate variables in our dataset in terms of representing the mean annual temperature (MAT) and mean annual total precipitation (MAP) of study site. The first set refers to the original study descriptions, and the second set is derived from the global climate dataset[11] with standardized 30-year reference periods. Also, to allocate data to ecosystem types we prepared a separate biome variable in the current dataset for more standardized representation of the compiled terrestrial ecosystem types. Similarly, the descriptions of soil texture in the original studies varied depending on the reference system used. Therefore, in those cases where the weight percentages of sand, silt, and clay where available, we standardized the available information to conform to the USDA texture triangle terminology (https://www.nrcs.usda.gov/resources/education-and-teaching-materials/soil-texture-calculator). The aim of this dataset publication is to provide gross rates of various N transformation processes for the study locations (see site map in Fig. 2) but also to delineate global patterns of functional relationships of N transformations depending on environmental conditions (i.e., soil, climate, ecosystem/biome). For the second part, this dataset includes a data from a robust machine learning (ML) modelling-based data imputation by utilizing and identifying relationships between closely associated N processes.

Methods

Data acquisition. To collect studies reporting measurements of soil gross N processes, Scopus and Web of Science (WoS) databases were searched in March 2022 using the following keywords: (TITLE-ABS-KEY (“gross N transformation*“ AND soil) OR TITLE-ABS-KEY (“^{15}N isotope dilution” AND soil) OR TITLE-ABS-KEY (“^{15}N tracing” AND soil) OR TITLE-ABS-KEY (Ntrace AND soil) OR TITLE-ABS-KEY (FLUAZ AND soil) OR TITLE-ABS-KEY (“gross nitrogen mineralization” AND soil) OR TITLE-ABS-KEY (“gross N immobilization”

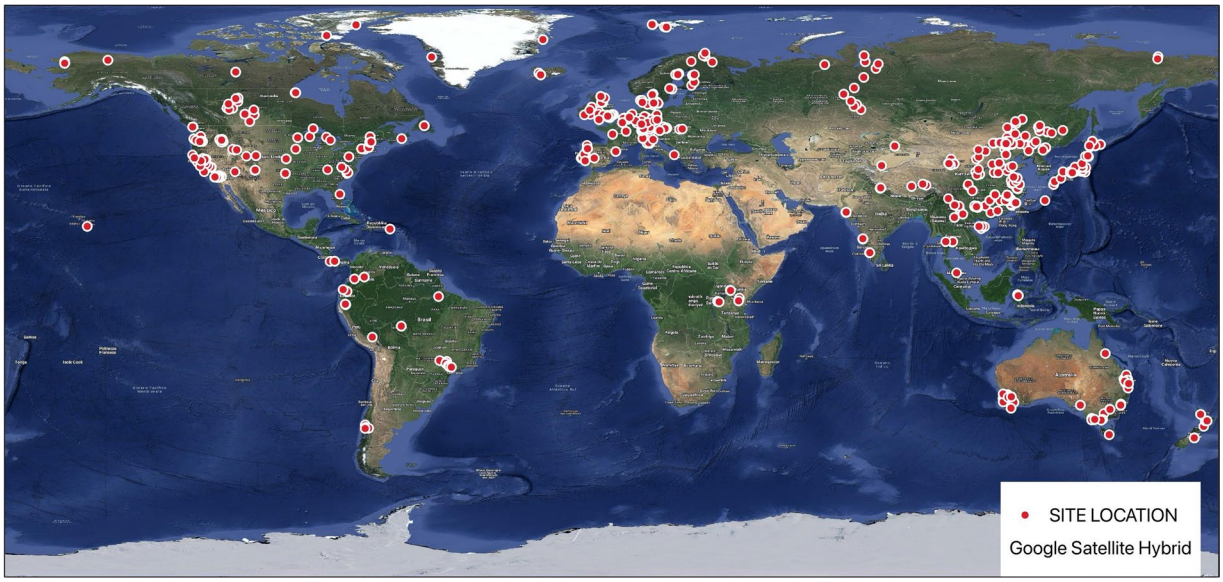

Fig. 2 Site map for the global scale dataset of gross N transformation rates.

AND soil) OR TITLE-ABS-KEY ("gross NH_4^+ immobilization" AND soil) OR TITLE-ABS-KEY ("gross NO_3^- immobilization" AND soil) OR TITLE-ABS-KEY ("gross nitrification" AND soil) AND NOT TITLE-ABS-KEY (ocean) AND NOT TITLE-ABS-KEY (marine) AND NOT TITLE-ABS-KEY (sea)) AND (LIMIT-TO (LANGUAGE, "English")). One of the search keywords, Ntrace, is a parameter estimation method that quantifies gross N transformation rates based on ^{15}N trace measurements and has been used in more than 200 peer-reviewed publications[4]. The search strategy returned 820 and 519 studies from Scopus and WoS respectively. Duplicate studies retrieved from the two databases were detected and removed based on Digital Object Identifier (DOI), or, in case of missing DOI, on title. After removing duplicates, we obtained a list of 820 studies that were screened by reading their abstracts and unsuitable papers were excluded.

The resulting 509 candidate studies were individually examined and included in the data compilation if, 1) soil N processes rates were estimated using ^{15}N isotope techniques 2) the measured rates or means were clearly reported in the text or tables or 3) could be retrieved from a graphical representation. Finally, using these selection criteria, we compiled a dataset of 1552 observations extracted from the final selected 581 sites as reported in 331 studies, of which 215 (65%) and 115 (35%) determined gross N process rates using the isotope dilution and N tracing approaches, respectively. Of the 331 studies, 291 (88%) carried out soil incubation in laboratory, while 40 (12%) were from *in situ* observations. Furthermore, keyword analysis was carried out by natural language processing models for the final collected titles and abstracts. All the title and abstract letters were converted to lowercase for a consistent data input to the model functions in R 'udpipe' package[12], which was aimed to examine the common goals of these gross rate measurements from wide geographic locations and diverse ecosystem types globally.

Data wrangling. We extracted as many environmental variables as possible from the retrieved studies, directly from the text, tables or from supplementary materials. When authors referred to previous studies for soil chemical-physical characteristics or for other variables of interest (e.g., pH, organic matter percentage, carbon content, nitrogen content, carton to nitrogen ratio, moisture content) we searched for the cited studies and extracted data from primary source references. In some cases, the corresponding authors of the study were contacted for data sharing request. To estimate the numerical means from data visualization and figures, we used the plot digitization tool WebPlotDigitizer (https://apps.automeris.io/wpd/).

Nitrogen process rates were expressed as mg N kg^{-1} soil day^{-1} but also as mg N g^{-1} carbon (C) day^{-1}. When data in the texts were reported as mg N m^{-2} day^{-1}, they were converted to mg N kg^{-1} day^{-1} using the soil bulk density and depth of soil sampling. If soil bulk density and sampling depth were not available in the text, data were not extracted. If the latitude and longitude information of original soil sampling was not clearly given in the original paper, the location coordinates of named sites were estimated from Google Earth map application. Measurements that represent different ecosystem types, plant species, and treatment levels within a single study were recorded as separate observations. The climate information of the study site, such as MAT (°C) and MAP (mm yr^{-1}) was recorded as mentioned in the article, or, if not reported, was extracted from the global climate database[11] using the location information (i.e. latitude and longitude) of the study site. When available, the measurement of ambient N deposition rates was extracted as well.

Values of soil water content during the experiment were extracted as well. However, for *in-situ* incubations soil water content was only occasionally reported. Moreover, soil water content was expressed using different metrics, that is gravimetric water content (GWC; g g^{-1}), percent of water-holding capacity (%WHC) or percent of water-filled pore space capacity (%WFPS), the latter a proxy of water and oxygen availability to soil microbes.

Name	Unit	Description
TN	g N kg^{-1} soil	Total soil nitrogen content in dry weight.
C_N		Soil total organic carbon to nitrogen mass ratio.
Ammonium	mg kg^{-1} soil	Extractable NH_4^+ concentration in the soil sample.
Nitrate	mg kg^{-1} soil	Extractable NO_3^- concentration in the soil sample.
MBN	mg kg^{-1} soil	Microbial biomass nitrogen content.
MBC		Microbial biomass carbon content.
MBC_N		Microbial biomass carbon to nitrogen ratio.
NNM	mg N kg^{-1} soil day^{-1}	Net nitrogen mineralization rate.
NNR	mg N kg^{-1} soil day^{-1}	Net nitrification rate of the soil.
GNM	mg N kg^{-1} soil day^{-1}	Gross nitrogen mineralization rate.
GNR	mg N kg^{-1} soil day^{-1}	Gross total nitrification rate.
GNRa	mg N kg^{-1} soil day^{-1}	Gross autotrophic nitrification rate.
GNRh	mg N kg^{-1} soil day^{-1}	Gross heterotrophic nitrification rate.
DNRA	mg N kg^{-1} soil day^{-1}	Dissimilatory nitrate reduction to ammonium.
INH_4	mg N kg^{-1} soil day^{-1}	Gross NH_4^+ immobilization rate.
INO_3	mg N kg^{-1} soil day^{-1}	Gross NO_3^- immobilization rate.
INN	mg N kg^{-1} soil day^{-1}	Gross NH_4^+ and NO_3^- immobilization rate.

Table 1. Nitrogen related variables in the dataset.

After the extraction of each metric, they were converted to %WFPS according to the information available in the studies. In some cases %WFPS was computed by dividing the volumetric water content (calculated as GWC*soil bulk density/water density) by total soil porosity, with the latter calculated according to soil porosity = 1 – (soil bulk density/2.65) assuming a soil particle density of 2.65 (g cm^{-3}) according to Linn and Doran[13]. In other cases, %WFPS was obtained by dividing %WHC values reported in the studies by 1.415, following Franzluebbers[14].

After the primary data extraction steps, standardized site information was appended taken into account discrepancies in the terminology of soil and climate classification systems. For example, regional studies adopted different systems for the classification of the soil type. Therefore, after the initial extraction of soil type, all soil types were coherently coded following the World Reference Base for Soil Resources (WRB) system[15]. Harmonization of soil classification (WRB IUSS) was performed on both local and regional classifications, where applicable, missing soil types were added. Either by cross-referencing the point location against regional soil maps if available, or the "best fitting" Soil Typological Unit (STU) classification reported in the Soil Map Unit of the Harmonized World Soil Database ver.1.1[16]. and ISRIC-WISE derived soil properties database[17]. Final soil classification was organized in the two main levels of WRB Reference Group (WRB_rsg code) and WRB First Qualifier (WRB_1qual code).

Similarly, we adopted the Köppen-Geiger systems for the climate classification of the study site location, including the extraction of 30-year average climate values based on a recently published spatial reference11. For the ecosystem types, we manually standardized the terms by defining commonly identified biomes as grasslands, croplands, forests, shrublands, desert, wetlands, and tundra (including 'polar' and 'alpine' tundra). Moreover, forest biomes were primarily coded into tropical, temperate, and boreal types, and secondarily into needleleaf, broadleaf, mixed broadleaf-needleleaf forests for 'leaf growth form' factor and into deciduous and evergreen forests for 'leaf longevity' factor. Also, forests were either known as natural 'Forest' or artificial 'Plantation' if such information was available.

Data imputation. Not all studies were designed to measure every type of N pathway known for the respective terrestrial ecosystem (see an example list of N pathway variables in Table 1). Among the N-related variables in Table 1, eight processes are representative of the gross N transformation rates measured in the studies, i.e., gross N mineralization (GNM), gross nitrification (GNR), gross autotrophic nitrification (GNRa), gross heterotrophic nitrification (GNRh), dissimilatory nitrate (NO_3^-) reduction to NH_4^+ (DNRA), immobilization of NH_4^+ (I_{NH4}), immobilization of NO_3^- (I_{NO3}), and immobilization of NO_3^- and NH_4^+ (INN).

Some N rate variables are not independent from each other, for example, INN ($I_{NH4} + I_{NO3}$) which was sometimes reported instead of the two subrates I_{NH4} and I_{NO3}. In general, most studies focused on a few or coupled N transformation pathways based on rate determining factors governed by the environment or by the experimental set ups. Consequently, the compiled dataset presents most of the gross rate variables but may include some rows where data are not reported. However, a significant number of complete data rows were available for the major variables (see Table 1), for example, about 50% of the total observations (774/1552) reported the gross rates of GNM, GNR, and INN, and this ratio went up to 75% for the cases with both GNM and GNR but not necessarily INN. More detailed diagnosis on the missing data is discussed in the technical validation section below.

In the compiled dataset, N related variables were analysed by correlating site environmental variables serving as possible predictors for the variability of gross N transformation rates. We used a machine learning data imputation on the original dataset. The imputation of the data was mainly aimed at more representative and commonly measured N pathways such as GNM and GNR, where a high probability of robust imputation was expected given the relatively low proportion of missing data rows compared to the other N rates. Caution is required for the direct use of imputation outcomes specifically for local scale interpretation of certain N pathways.

The R 'missRanger' package was used for the imputation[18]. Its main function fits data into a random forest (RF) algorithm and makes predictions on the missing values, available for both categorial and numerical. All existing values are potentially predictors for the missing value at a given data row, while the RF model is learning the relationships between variables. The function iterates the RF modelling while generating prediction of each missing value across all variables and evaluating the predicted values by the average of out-of-bag errors for each variable. The iteration stops when the error metrics on averages does not improve in the subsequent modelling, and the final model outcome includes the imputed values from the best iteration hitherto. The imputation results from ten independent missRanger runs were summarized by mean and standard deviation for each imputed value. The primary key (ID) is the same as the original dataset so that the imputed results can be associated with the original site variables.

Data Records

Original compilation. The original compiled data and imputed version are deposited in the figshare repository[19] under the creative commons license CC BY 4.0 (deposition in preparation). It includes 1552 observations with partially missing gross N process rates for most observations. The number of complete observations for the compiled eight individual N processes (i.e., GNM, GNR, GNRa, GNRh, INN, I_{NH4}, I_{NO3}, and DNRA) is 269 or for the three more commonly measured processes (i.e., GNM, GNR, and INN) is 774. The environmental variables that explain the site conditions for each observation vary depending on the site type (e.g., forest ecosystem has additional variables for the forest types), but complete for the standardized ecosystem types (biome) as in seven representative biomes, harmonized soil classes, and climate variables.

The dataset upload was done as a single Microsoft Excel file including five spreadsheets, and has a relational structure, i.e. they share a common column (the ID) so data can be connected among the tables. The first is a README file that contains comments. The second spreadsheet includes Metadata information with descriptions of each compiled variable such as data type, acquisition process, and units if applicable. The third, contains Metadata of the imputed data. The fourth spreadsheet includes the compiled dataset of original records with completed location coordinates and additional environmental variables appended in this work (Table 2).

Imputation results. The fifth spreadsheet in the excel file is a table from machine learning (ML) data imputation outcomes[19]. Among the N process related variables (Table 1), the variables to be updated by the ML data imputation were determined based on the proportion of missing data rows (Table 2) less than 50%. However, GNRh and GNRa were included in the table expecting high dependence on the mostly available GNR data. The resulting values are the mean of the best iteration outcome from each independent missRanger model run after 10 replicates, with standard deviations for those mean values[19]. The prediction of missing data was made by a RF algorithm with 1000 trees on the site environmental variables. Specifically, the imputation model excluded variables regarding the publication information and those descriptive non-categorical variables, such as original soil class descriptions from the publication (i.e., Soil_class) and dominant plant descriptions (i.e., Plant_dominant) as shown in Table 2 as variables in character type data with more than 100 unique counts.

While the data imputation was aimed at missing N rate values, the RF-based predictions were regardless performed on all the missing values of any input variables. Thus, the complete observations were used as primary explanations between the environmental variables, and then less complete observations were predicted by relatively more complete observations. The model performance metrics were suggested by out-of-bag errors of all the prediction results, regardless of environmental or gross N rate variables, and on average 0.83 for the ten best prediction models (selected each from ten random replicate missRanger() runs in this case), but we also report the individual R-squared value for the imputed N rates in the Metadata table in the dataset[19]. Note that ID are the same across the two spreadsheets so that the imputed results can be appended with other environmental variables in the original compilation.

Technical Validation

Data extraction. While extracting data from the previous studies, we explored the research relevance of the final literature collection for the gross N rate and related environmental data extraction. Frequent keywords recognized from the titles were soil types and microbial related terms (Fig. 3), which shows the importance of soils in N transformation pathways and supports our aim to represent various terrestrial ecosystem types through this global dataset work. Interestingly, frequent keywords recognized from the abstract texts were related to the measurement techniques for gross rates, i.e. isotope dilution and dilution technique that, if merged, represent the most common term reported in the abstract keywords (i.e., isotope dilution technique). This probably highlights the research interest in clarifying the method used for the gross rates. Other common terms are related to soil organic matter, i.e. organic matter, organic C and organic layer, and greenhouse gas emissions, i.e. nitrous oxide (N_2O). Yet, abstract keywords are also characterized by the presence of terms related to the microbial ecology of the N cycle, i.e. functional gene, microbial community, and community structure (Fig. 3). While these studies generally recognize the central role of microorganisms in the N transformation processes, actual bulk and molecular microbiology data or related measurements are not as commonly performed as we hoped for. As a result, a systematic inclusion of data related to the microbial components was deemed not possible at this stage. Where possible, we attempted to at least provide microbial biomass information, although even this information is largely missing from the final dataset (Table 2). We pursued a complete dataset with as many rate variables as possible for each observation for understanding gross N transformations in the terrestrial ecosystems.

Data imputation. Originally, the data imputation was aimed to improve the global data coverage for various N transformation processes. For the ML model training, the input data included most of the site environmental

Variable	Data type	Missing count	Missing (%)	Unique count
ID	numeric	0	0.0	1552
Authors*	character	0	0.0	205
Year	numeric	0	0.0	31
Publication	character	0	0.0	331
Journal	character	0	0.0	75
DOI	character	17	1.1	326
LONDD	numeric	0	0.0	581
LATDD	numeric	0	0.0	603
Climate	character	1067	68.8	58
Elevation	numeric	1020	65.7	150
MAP	numeric	376	24.2	286
MAT	numeric	525	33.8	173
KC18_MAP	numeric	0	0.0	491
KC18_MAT	numeric	0	0.0	527
KC18_main	character	0	0.0	5
KC18_sub	character	0	0.0	23
Ecosystem	character	6	0.4	8
Plant_dominant	character	386	24.9	382
Biome	character	0	0.0	9
Bio_leaf_grow	character	861	55.5	16
Bio_leaf_long	character	876	56.4	16
Leaf_grow	character	861	55.5	6
Leaf_long	character	876	56.4	6
Study_type	character	0	0.0	2
Ambient_N	numeric	1306	84.1	66
N_fertilized	logical	0	0.0	2
Soil_class	character	398	25.6	271
WRB_rsg	character	16	1.0	30
WRB_1qual	character	53	3.4	73
Soil_horiz	character	0	0.0	3
Top_cm	numeric	84	5.4	19
Bottom_cm	numeric	84	5.4	40
Soil_layer	character	82	5.3	3
Clay_orig	numeric	899	57.9	54
Silt_orig	numeric	945	60.9	75
Sand_orig	numeric	937	60.4	88
Soil_texture_orig	character	709	45.7	27
Clay_perc	numeric	945	60.9	52
Silt_perc	numeric	945	60.9	78
Sand_perc	numeric	945	60.9	90
Soil_texture_class	character	669	43.1	13
WHC	numeric	898	57.9	35
WFPS	numeric	362	23.3	87
Soil_pH	numeric	0	0.0	67
Soil_pH_class	character	0	0.0	9
TOC	numeric	138	8.9	597
TN	numeric	235	15.1	180
C_N	numeric	213	13.7	259
Ammonium	numeric	650	41.9	261
Nitrate	numeric	651	41.9	305
MBC	numeric	1201	77.4	297
MBN	numeric	1150	74.1	229
MBC_N	numeric	1286	82.9	134
NNM	numeric	1171	75.5	287
NNR	numeric	1204	77.6	248
GNM	numeric	105	6.8	818

Continued

Variable	Data type	Missing count	Missing (%)	Unique count
GNR	numeric	286	18.4	612
GNRa	numeric	1191	76.7	260
GNRh	numeric	1207	77.8	127
DNRA	numeric	1157	74.5	100
INH_4	numeric	658	42.4	522
INO_3	numeric	744	47.9	286
INN	numeric	755	48.6	533

Table 2. Diagnosis of all variables with missing data in the global dataset. This summary table was created by a data diagnostic function in R 'dlookr' package[26]. *Variable was filled by the last name of the first author, followed by '*et al.*' if more than or equal to three authors, or the last names of both authors for two author papers. Note that some last names are in the exact same spelling. Please refer to the metadata of the dataset for more variable descriptions[19].

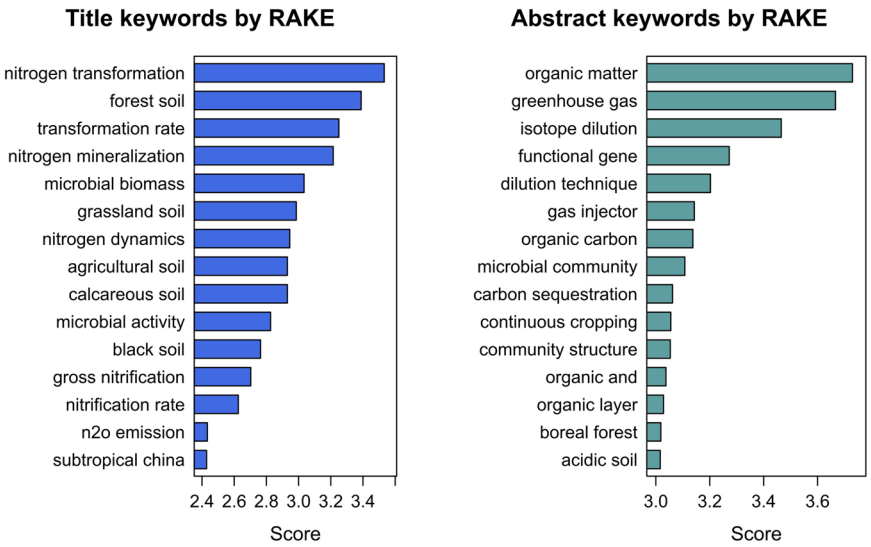

Fig. 3 Keyword analysis results for the study titles and abstracts. The analysis was performed by RAKE (Rapid Automatic Keyword Extraction) algorithm for natural language processing (NLP), the function offered by R 'udpipe' package. The study titles and abstracts were collected from the final literature selection for data extraction in this work. Converted all text to lowercase for consistency in the NLP keyword detection.

variables as potential predictors for the targeted N process rates. Also, it included all the N process related variables (Table 1), meaning that the imputation target variables served also as predictors while their missing values were predicted. As such, the reason why this ML modelling did not test or consider the independency of input variables was because the goal was to generate most robust imputed values based on available data and non-considered relationships between variables, rather than to examine the importance of environmental factors on specific N pathways. As a result, the predicted and filled N process rates followed similar distribution of the original values as shown in Fig. 4 for the three major processes. Overall, the filling was biased for the mid-range values, avoiding unrealistic values from outside of the original distribution but following the frequency of existing values, which further emphasizes the need for future data collection targeted at filling current missing values including detailed environmental parameters.

As such, some of the N related variables were expected to be dependent on each other, producing substrates in soil environments for the subsequent process by microbial activities, and thus their gross rates were expected to be correlated (Fig. 5, left). This correlation was deemed to help the ML model's pattern learning, aimed for robust data-driven prediction performance. The predictor variable importance for a certain N pathway and gross rate variation can be explored by future data analysis studies. In this regard, any direct use of the imputed rate values for a specific study site was not intended here, but a comparison of different N pathways can refer to summary statistics for a regional or continental scale study or by ecosystem type (e.g., parameter inputs to a global biogeochemical cycle model).

Thus, our ML prediction-based imputation results suggest the need of future simultaneous gross rate measurements to provide an empirical basis of theoretically coupled N pathways by a substrate-product relationship. For example, INN is well correlated with the measured GNM, which agree with the mineralization-immobilization turnover (MIT) model according to which there is a continuous transfer of mineralised N into microbial biomass and vice versa. In such coupled pathways, GNM technically produces NH_4^+

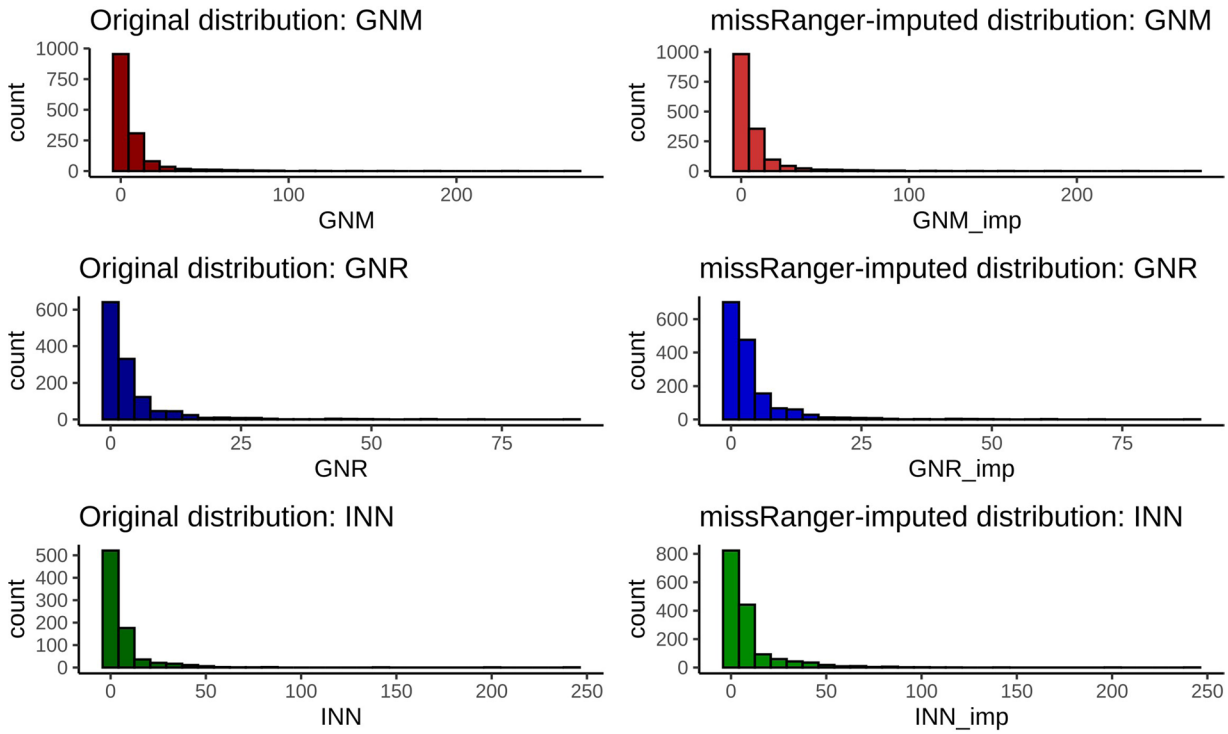

Fig. 4 Frequency count for the ranges of gross N rates per representative three processes.

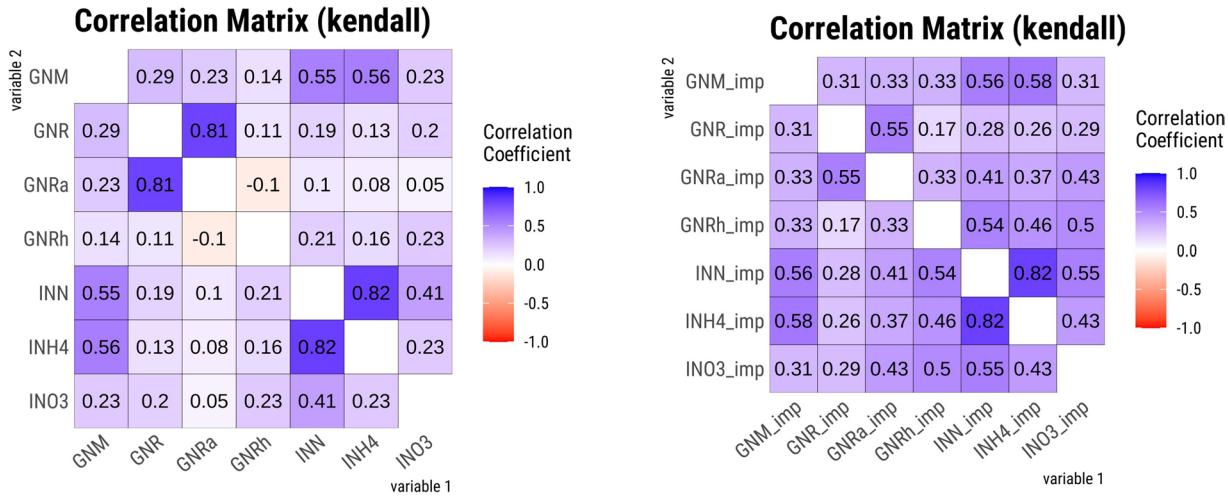

Fig. 5 Correlation matrix for the gross N rate variables with the values as compiled (left) or those imputed (right). Correlation coefficients by pairwise complete observations on each pair of variables (left) as default setting in the correlate function in R 'dlookr' package[26]. The correlation coefficients are shown for the imputation results not including the observed values (e.g., the case count for GNM_imp is 105 which is the same number of missing counts as in Table 2).

and thus is better correlated with I_{NH4}, compared to that of I_{NO3} as shown by the difference in correlation coefficients. Another example of the paired N transformation pathways is that the GNRa that has been identified as the main pathway for GNR process[20] results in high correlation coefficients in the original data compilation. In contrast, GNRh is likely a secondary player for the bioavailable N production in soils[21]. However, it should be noted that in studies where the ^{15}N dilution technique has been applied, the measured total gross NO_3^- production includes both autotrophic and heterotrophic NO_3^- production[22]. Furthermore, recent studies suggest that plants can stimulate heterotrophic nitrification, therefore the fact that heterotrophic nitrification does not seem to be an important process, may stem from the absence of plants during soil incubations[23] (see also below).

The imputed results may correspond to the existing dependences between paired N pathways, but only up to a certain degree, which is theoretically supported as described above. The correlation coefficients for the

imputed values were overall exaggerated (Fig. 4 right). Still, a relatively low initial dependence between GNM and GNR in the original dataset was preserved for the imputed values, in part attributable to the initially low fraction of the missing counts (Table 2). It also suggests that the imputation of both variables may have resulted from the predictions of other site environmental variables, although this aspect is not explored in detail which is not in the scope of this dataset work. Other weakly dependent gross N rates are recommended to be further explored in future studies, whether there could be an explanation based on environmental control or unrealized dependence due to a relative lack of measurement data. For example, the GNRh has been studied for potentially tight correlations with I_{NO3} and GNM[6–8], which was taken into account by the imputation of missing rates.

Usage Notes

The presented dataset has some limitations that warrant consideration. The addition of ^{15}N can stimulate certain processes, and thus their measurements might be overestimated. Moreover, in soil incubations conducted under optimal conditions, such as laboratory studies, the determined rates could also be overestimated and then should be considered as potential rather than actual rates[24]. Soil N transformations are mediated by microorganisms whose activity is influenced by plants either through the release of root exudates or through N uptake[25]. Nevertheless, measured gross rates presented in the dataset were obtained in the absence of plants, which suggests that caution is needed in interpreting the presented data. However, despite these limitations, the present dataset offers a unique opportunity to enhance our mechanistic understanding of the global N cycle. Lastly, despite the experimental sites are globally distributed, some continents or regions of the world are less represented in our dataset. It is the case of Africa for which a future increase in the number of studies is desirable.

The dataset is available as an easy to access spreadsheet format, aiming to provide the scientific community with an overview of the global availability of existing measurements to date. For each data point included, we provide detailed source information, hence, researchers will be able to refer to the original article and apply filters specifically tailored to their analyses. Also, new data can easily be added by referring to the metadata records in the dataset as well as the method described in this paper. The imputation outcome is subjective to future updates as new records or variables are added. The annotated R code scripts to reproduce all the figures in this paper and to perform the machine learning data imputation as described in the above section are available and encouraged to be modified for the purpose of data analysis. The specific R packages used for the modelling and figure production are cited throughout this descriptor and should be installed as guided by the developers to properly run the provided R codes. Readers are encouraged to refer to the detailed specs for each package and functions through the package vignettes archived in CRAN network.

Code availability

The R code scripts, and source data tables are found together in the dataset upload in the figshare repository[19]. Please follow the instructions in the README text file for details.

Received: 24 April 2024; Accepted: 5 September 2024;
Published online: 19 September 2024

References

1. Vitousek, P. M. & Howarth, R. W. Nitrogen limitation on land and in the sea: How can it occur? *Biogeochemistry* **13**(2), 87–115 (1991).
2. Hart, S. C. *et al*. Dynamics of gross nitrogen transformations in an old-growth forest: The carbon connection. *Ecology* **75**(4), 880–891 (1994).
3. Kirkham, D. & Bartholomew, W. V. Equations for following nutrient transformations in soil, utilizing tracer data. *Soil Science Society of America Proceedings* **18**(1), 33–34 (1954).
4. Jansen-Willems, A. B., Zawallich, J. & Müller, C. Advanced tool for analysing ^{15}N tracing data. *Soil Biology & Biochemistry* **165**, 108532 (2022).
5. Stark, J. M. & Hart, S. C. High rates of nitrification and nitrate turnover in undisturbed coniferous forests. *Nature* **385**(6611), 61–64 (1997).
6. Elrys, A. *et al*. Patterns and drivers of global gross nitrogen mineralization in soils. *Global Change Biology* **27**(22), 5950–5962 (2021).
7. Elrys, A. *et al*. Global gross nitrification rates are dominantly driven by soil carbon-to-nitrogen stoichiometry and total nitrogen. *Global Change Biology* **27**(24), 6512–6524 (2021).
8. Elrys, A. S. *et al*. Global patterns of soil gross immobilization of ammonium and nitrate in terrestrial ecosystems. *Global Change Biology* **28**(14), 4472–4488 (2022).
9. Elrys, A. S. *et al*. Integrative knowledge-based nitrogen management practices can provide positive effects on ecosystem nitrogen retention for sustainable agriculture. *Nature Food* **4**, 1075–1089 (2023).
10. Booth, M. S., Stark, J. M. & Rastetter, E. B. Controls on nitrogen cycling in terrestrial ecosystems: a synthetic analysis of literature data. *Ecological Monographs* **75**(2), 139–157 (2005).
11. Cui, D. Y. *et al*. A 1 km global dataset of historical (1979-2013) and future (2020-2100) Koppen-Geiger climate classification and bioclimatic variables. *Earth System Science Data* **13**(11), 5087–5114 (2021).
12. Wijffels, J., *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. 2023.
13. Linn, D. M. & Doran, J. W. Effect of water-filled pore space on carbon dioxide and nitrous oxide production in tilled and nontilled soils. *Soil Science Society of America Journal* **48**(6), 1267–1272 (1984).
14. Franzluebbers, A. J. Holding water with capacity to target porosity. *Agricultural & Environmental Letters* **5**(1), e20029 (2020).
15. WRB, I. W. G. World Reference Base for Soil Resources 2014, update 2015 International soil classification system for naming soils and creating legends for soil maps. *World Soil Resources Reports No. 106*. (FAO, Rome, Italy, 2015).
16. Nachtergaele, F., H. Van Velthuizen, and L. Verelst, *Harmonized World Soil Database Version 1.1*. 2009, FAO/IIASA/ISRIC/ISS-CAS/JRC: FAO, Rome, Italy and IIASA, Laxenburg, Austria.
17. Batjes, N. H., *ISRIC-WISE derived soil properties on a 5 by 5 arc-minutes global grid (ver. 1.2.)*, in *ISRIC - World Soil Information*. 2012: Wageningen.
18. Mayer, M. *missRanger: Fast Imputation of Missing Values*. 2023.

19. Byun, E. *et al. A global dataset of gross nitrogen transformation rates across terrestrial ecosystems. figshare* https://doi.org/10.6084/m9.figshare.26886070 (2024).
20. Faeflen, S. J. *et al.* Autotrophic and heterotrophic nitrification in a highly acidic subtropical pine forest soil. *Pedosphere* **26**(6), 904–910 (2016).
21. Martikainen, P. J., Heterotrophic nitrification – An eternal mystery in the nitrogen cycle. Soil Biology & Biochemistry, 2022. **168**.
22. Barraclough, D. & Puri, G. The use of ^{15}N pool dilution and enrichment to separate the heterotrophic and autotrophic pathways of nitrification. *Soil Biology & Biochemistry* **27**(1), 17–22 (1995).
23. He, X. *et al.* ^{15}N tracing studies including plant N uptake processes provide new insights on gross N transformations in soil-plant systems. *Soil Biology & Biochemistry* **141**, 107666 (2020).
24. Davidson, E., *Fluxes of nitrous oxide and nitric oxide from terrestrial ecosystems*, W.B.W. J.E Rogers, Editor. 1991, American Society for Microbiology: Washington. p. 219–235.
25. Nardi, P. *et al.* Biological nitrification inhibition in the rhizosphere: determining interactions and impact on microbially mediated processes and potential applications. *FEMS Microbiology Reviews* **44**(6), 874–908 (2020).
26. Ryu, C. *{dlookr}: Tools for Data Diagnosis, Exploration, Transformation*. (2024).

Acknowledgements

The research was partially supported by Yonsei University Research Fund (2023-22-0433) and by the National Research Foundation of Korea grant (2023-11-0917). J.I.A. was supported by NSF award (grant number: 2133863).

Author contributions

E.B. conceptualization and initial draft of the manuscript; dataset review and update. C.M. reviewed and edited manuscript. B.P. contributed to the data compilation; prepared climate variables. R.N. structure of the alphanumeric and spatial dataset; harmonization of soil classification. F.R. reviewed and edited manuscript. P.V.C. reviewed and edited manuscript. J.-B.Z. reviewed and edited manuscript. G.M. reviewed and edited manuscript. A.B.J.-W. reviewed and edited the manuscript. W.H.Y. provided original data and reviewed the manuscript. R.U. provided original data and reviewed manuscript. J.I.A. helped to organize the dataset structure and writing of the manuscript. U.N. contributed to the data compilation. A.S.E. reviewed and edited manuscript. P.N. conceptualization and initial draft of the manuscript; reviewed and edited manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.