https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

A robustness diagnostic framework for NMIP ensembles: Application to NMIP2 soil N_2O emission estimates

Motoko Inatomi[1]

[1]Institute for Agro-Environmental Science, NARO, Tsukuba, 305-8604, Japan

Correspondence to: Motoko Inatomi (inatomi.motoko312@naro.go.jp)

Abstract. Multi-model ensembles, such as the global N_2O Model Intercomparison Project (NMIP), are essential for quantifying terrestrial nitrous oxide (N_2O) fluxes; however, interpreting where ensemble results are reliable and where they are not remains challenging. Here, we propose a robustness diagnostic framework (RDF) that classifies each grid cell into four categories: Robust-increase, Robust-decrease, Divergent and Uncertain, based on the three metrics of ensemble-mean change, model agreement on the sign of change and inter-model standard deviation. By explicitly separating directional disagreement (Divergent) from quantitative disagreement (Uncertain), the RDF reveals the nature of model uncertainty rather than its magnitude alone. We applied the RDF to soil N_2O emission estimates from eight terrestrial biosphere models participating in NMIP phase 2 (NMIP2), comparing pre-industrial (1850s) and contemporary (2010s) periods. Of 56,852 valid grid cells, 40.8% were classified as Robust-increase, 2.6% as Robust-decrease, 36.1% as Divergent and 20.5% as Uncertain. Stratification by land-use type revealed that the nature of uncertainty differed fundamentally: cropland-dominated regions were dominated by Uncertain (72.6%), indicating an agreement on the direction of N_2O increase but a large quantitative spread, whereas forest-dominated regions were dominated by Divergent (50.0%), indicating disagreement on the direction of change itself. Pasture-dominated regions exhibited the highest robustness (59.1% Robust-increase). The inter-model spread correlated strongly with nitrogen input intensity (Spearman $\rho = 0.75$), and the Divergent to Uncertain transition followed a gradient of cropland fraction. These contrasting patterns implied that different model improvement strategies were needed: observational benchmarking of emission magnitudes for croplands, and improved process understanding of the competition between carbon dioxide (CO_2) fertilization and warming effects in forests. The proposed framework is general and applicable to any multi-model ensemble of biogeochemical change.

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

1 Introduction

Nitrous oxide (N_2O) is a potent greenhouse gas with a 100-year global warming potential of 273 (Foster et al., 2021) and is the dominant anthropogenic ozone-depleting substance emitted in the 21st century (Ravishankara et al., 2009). Its atmospheric mole fraction has increased from approximately 270 ppb in 1750 to 336 ppb in 2022 (WMO, 2023), with agricultural soils representing the largest anthropogenic source through synthetic nitrogen (N) fertilizer application and manure management (Davidson, 2009; IPCC, 2019)

Process-based terrestrial biosphere models (TBMs) are essential tools for estimating soil N_2O fluxes and understanding their responses to environmental change. The global N_2O Model Intercomparison Project (NMIP) coordinates multi-model simulations of terrestrial N_2O emission using consistent driving datasets, enabling quantification of the structural variability arising from different model formulations of key biogeochemical processes, such as nitrification and denitrification (Tian et al., 2018). The second phase of NMIP (NMIP2) comprises eight TBMs and covers the period 1850−2020, representing the most comprehensive process-based ensemble of terrestrial soil N_2O estimates to date (Tian et al., 2024).

Recent studies have utilized NMIP2 outputs to advance our understanding of N_2O emissions. For example, Li et al. (2024) developed a hybrid machine learning framework that emulates the ensemble-mean behaviour of NMIP2 models to project emission factors (EFs) under future climate and management scenarios. While their approach effectively captures the mean response to environmental drivers, it does not address inter-model disagreement, which is the focus of the present study. Previous studies have primarily focused on estimating emission magnitudes or projecting future trends. However, a systematic diagnosis of where and how NMIP2 models agree or diverge, and how the nature of that disagreement varies across land-use types, has not been conducted. Such a diagnosis is a prerequisite for identifying where observational constraints and process improvements are most needed.

In multi-model ensemble studies, inter-model uncertainty has conventionally been characterized using the ensemble mean ± standard deviation (e.g. Tian et al., 2020, 2024). While this approach quantifies the overall spread, it does not distinguish between two qualitatively different situations: regions where models disagree on the direction of change, and regions where models agree on the direction but differ in the magnitude. Analogous challenges have been addressed in climate projection studies. Knutti and Sedláček (2013) developed a robustness metric for CMIP5 projections that separately identifies regions of high model agreement, indicating the reliability of climate projections. Their approach was adopted in IPCC AR5 (Collins et al., 2013) and demonstrated that distinguishing between lack of signal and lack of agreement fundamentally improves the interpretation of multi-model results. While individual model maps and ensemble statistics are routinely reported for NMIP, these outputs do not readily answer the practical question of where model results can be trusted and where they cannot. A classification system enabling any user, whether a modeller, observer, or policymaker, to readily identify the nature of model agreement at each location has been lacking for biogeochemical model ensembles.

Here, we propose a diagnostic framework of estimation robustness (hereafter RDF) that integrates the three complementary metrics of change magnitude, model agreement on the sign of change and inter-model standard deviation, into a four-class

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

classification: Robust-increase, Robust-decrease, Divergent and Uncertain. This RDF explicitly separates directional disagreement from quantitative uncertainty, a distinction that implies fundamentally different strategies for model improvement. We applied the RDF to NMIP2 soil N_2O estimates, comparing pre-industrial (1850s) and contemporary (2010s) periods, and classified the results by land-use type using the Land-Use Harmonization 2 (LUH2) dataset (Hurtt et al., 2020) and by nitrogen input intensity using the HaNi dataset (Tian et al., 2022). Our objectives were to:

1. Propose a diagnostic framework of estimation robustness applicable to multi-model ensembles of biogeochemical change;
2. Apply the RDF to NMIP2 soil N_2O emissions (1850s vs. 2010s);
3. Quantify how the nature of model uncertainty varies across land-use types; and
4. Identify differentiated model improvement priorities for croplands and forests.

2 Data and methods

2.1 NMIP2 model ensemble and experiment

We used gridded soil N_2O emission estimates from NMIP2 (Tian et al., 2024). NMIP2 comprises eight process-based terrestrial biosphere models: CLASSIC, DLEM, ELM, ISAM, LPX-Bern, O-CN, ORCHIDEE and VISIT. All models were driven by consistent input datasets, including climate (CRU-JRA2.2), atmospheric carbon dioxide (CO_2) concentration, land cover change (LUH2; Hurtt et al., 2020), atmospheric nitrogen deposition, mineral nitrogen fertilization and manure nitrogen common driving data. The models differed substantially in their representation of key nitrogen cycling processes such as nitrification, denitrification and nitrogen fixation (Tian et al., 2024, Table S1).

For demonstration, we analysed the outputs from the SH1 experiment, in which all driving factors varied over the period 1850−2020. The SH1 results represent the "best estimates" of soil N_2O emissions because they include the combined effects of all factors that each model can account for (Tian et al., 2024). All data are provided as NetCDF files at 0.5° × 0.5° spatial resolution in units of g N m^{-2} yr^{-1}.

For each model m and each grid cell, we computed the decadal-mean emissions for the 1850s (1850−1859) and 2010s (2010−2019), and defined the change as:

$$\Delta_m = F_{2010s,m} - F_{1850s,m} \quad (1)$$

where F denotes the decadal-mean soil N_2O emission flux. The global mean F_{2010s} ranged from 0.053 (O-CN) to 0.083 (ELM) g N m^{-2} yr^{-1} across models (Table S1). We used Δ rather than absolute emission rates because 1850s baseline emissions differed substantially across models (0.026 to 0.057 g N m^{-2} yr^{-1}; Table S1), and Δ reduced the influence of differences in pre-industrial baseline emissions across models.

A grid cell was included in the analysis only if all eight models provided valid (non-missing) data for both the 1850s and 2010s, yielding 56,852 valid grid cells (21.9% of the global 0.5° grid).

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

2.2 Ancillary datasets

2.2.1 Land-use data

We used the Land-Use Harmonization 2 (LUH2; Hurtt et al., 2020) dataset to characterize the dominant land-use type of each grid cell. LUH2 provides annual land-use state fractions at 0.25° resolution. We aggregated these to 0.5° using 2 ×2 area-weighted averaging to match the NMIP2 grid resolution. For 2015 (representative of the 2010s), we extracted three land-use fractions: cropland (sum of c3ann, c4ann, c3per, c4per and c3nfx), forest (primf + secdf) and pasture (pastr + range). Because NMIP2 public outputs do not include agricultural metadata (e.g. crop types, fertilizer application rates, or management practices), we used the LUH2 cropland fraction as a proxy for agricultural intensity in our analysis.

2.2.2 Nitrogen input data

We used the HaNi dataset (Tian et al., 2022) to examine the relationship between anthropogenic nitrogen inputs and inter-model uncertainty. HaNi provides gridded nitrogen input estimates from 1860 to 2019 at 5 arcmin resolution. In this study, we used the 0.5°-regridded dataset prepared for NMIP2 model simulations, which ensured spatial consistency with NMIP2 outputs. Total nitrogen input was defined as the sum of synthetic nitrogen fertilizer (NFER), manure nitrogen (MANN) and atmospheric nitrogen deposition (NDEP) for the year 2015 (g N m^{-2} yr^{-1}).

2.3 Robustness diagnostic framework

We propose a diagnostic framework that classifies the estimate at each grid cell into one of four robustness categories based on three metrics computed from the eight-model ensemble of Δ values.

2.3.1 Diagnostic metrics

For each grid cell with $M = 8$ valid models, we computed:

(a) Ensemble-mean change:

$$\overline{\Delta} = \frac{1}{M}\sum_{m=1}^{M}\Delta_m \tag{2}$$

(b) Inter-model standard deviation:

$$SD(\Delta) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\Delta_m - \overline{\Delta}\right)^2} \tag{3}$$

(c) Agreement on the sign of change (κ_A):

$$\kappa_A = \frac{max(n_+, M - n_+)}{M} \tag{4}$$

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

where n_+ is the number of models with $\Delta_m > 0$. The agreement κ_A takes values from 0.500 (equal split, four vs. four) to 1.000 (all eight models agree). Grid cells with $\Delta_m = 0$ were excluded from the sign count; in practice, this occurred in negligibly few cells.

2.3.2 Four-class classification

The three metrics were used to classify the nature of model estimation using a hierarchical classification scheme (Fig. 1):

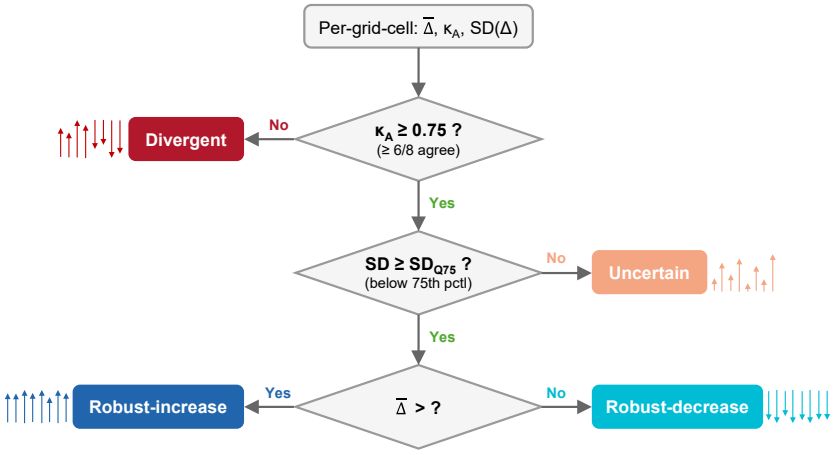

Figure 1. Flowchart of the robustness classification framework. Each grid cell was classified into one of four categories based on three diagnostic metrics: ensemble-mean change ($\overline{\Delta}$), model agreement on the sign of change (κ_A) and inter-model standard deviation of change (SD(Δ)). The schematic icons beside each class illustrate the model's behaviour. Arrows represent individual model responses, with their direction indicating the sign of change and their length indicating the magnitude.

1. Divergent ($\kappa_A < 0.75$): Fewer than 75% (i.e. six of eight) models agree on the direction of change, indicating directional disagreement.
2. Uncertain ($\kappa_A \geq 0.75$ and $SD(\Delta) \geq SD_{Q75}$): Models agree on the direction but exhibit a large quantitative spread.
3. Robust-increase ($\kappa_A \geq 0.75$, $SD(\Delta) < SD_{Q75}$ and $\overline{\Delta} > 0$): Models agree on an increase with a small spread.
4. Robust-decrease ($\kappa_A \geq 0.75$, $SD(\Delta) < SD_{Q75}$ and $\overline{\Delta} \leq 0$): Models agree on a decrease with a small spread.

The agreement threshold $\kappa_A = 0.75$ was selected because it corresponded to at least six of eight models agreeing on the sign of change, following the convention used in IPCC AR5 (Collins et al., 2013) and Knutti and Sedláček (2013). The SD threshold SD_{Q75} was defined as the 75th percentile of SD(Δ) across all 56,852 valid grid cells ($SD_{Q75} = 0.0392$ g N m^{-2} yr^{-1}), providing a data-driven separation between low and high inter-model spread. To test the sensitivity of the classification, we compared the results using an alternative agreement threshold of $\kappa_A \geq 0.875$ (seven of eight models; Table S2).

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

2.3.3 Design rationale

This RDF builds on the approaches used for climate model ensembles by Tebaldi et al. (2011), who proposed quantifying grid-cell-level model agreement, and by Knutti and Sedláček (2013), who introduced the stippling/hatching convention for CMIP5 projections that combines signal-to-noise ratio and model agreement. Our RDF extends these approaches by explicitly separating directional disagreement (Divergent) from quantitative disagreement (Uncertain), a distinction that is not captured by the conventional ensemble mean ± standard deviation approach. Moreover, this separation enables land-use-specific diagnosis of model uncertainty, which is essential for prioritizing model improvements.

2.4 Land-use stratification

To examine how the nature of model uncertainty varies with land-use type, we stratified grid cells into three categories based on LUH2 fractions:

- Cropland-dominated: cropland fraction ≥ 0.3
- Forest-dominated: forest fraction ≥ 0.5
- Pasture-dominated: pasture fraction ≥ 0.3

The asymmetric threshold reflects differences in the mechanisms through which each land-use type influences N_2O emissions. Croplands can significantly affect grid-cell N_2O emissions even at moderate coverage, mainly because of the large nitrogen inputs associated with fertilization; hence, a relatively low threshold of 0.3 is used. For forests, where N_2O dynamics are driven primarily by natural nitrogen cycling processes, we require a majority fraction (≥ 0.5) to ensure that the grid cell is indeed forest-dominated. Pastures receive intermediate nitrogen inputs, and a threshold of 0.3 is applied. These categories are not mutually exclusive; overlapping grid cells are retained in both categories. The sensitivity to alternative thresholds (cropland: 0.2, 0.5; forest: 0.3) is reported in Table S2.

For cropland-dominated grid cells, we further examined the progressive shift in robustness classification along a gradient of cropland fraction, using bins of 0.2 intervals (0.0−0.2, 0.2−0.4, 0.4−0.6, 0.6−0.8, 0.8−1.0).

2.5 Nitrogen input analysis

To investigate the relationship between anthropogenic nitrogen input intensity and inter-model uncertainty, we stratified all valid grid cells into quintiles (Q1−Q5) based on the total HaNi nitrogen input for 2015, where each quintile contains approximately equal numbers of grid cells (n ≈ 11,370). For each quintile, we computed the median SD(Δ) and the proportion of grid cells classified as Divergent and Uncertain. The monotonic association between nitrogen input and SD(Δ) was quantified using the Spearman rank correlation coefficient (ρ).

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

3 Results

3.1 Global robustness classification

The proposed methodology successfully captured the nature of multi-model estimates. The robustness classification of the 56,852 valid grid cells yielded the following global distribution (Fig. 2a): Robust-increase 23,192 (40.8%), Robust-decrease 1,490 (2.6%), Divergent 20,526 (36.1%) and Uncertain 11,644 (20.5%). The agreement threshold SD_{Q75} was 0.0392 g N m^{-2} yr^{-1}.

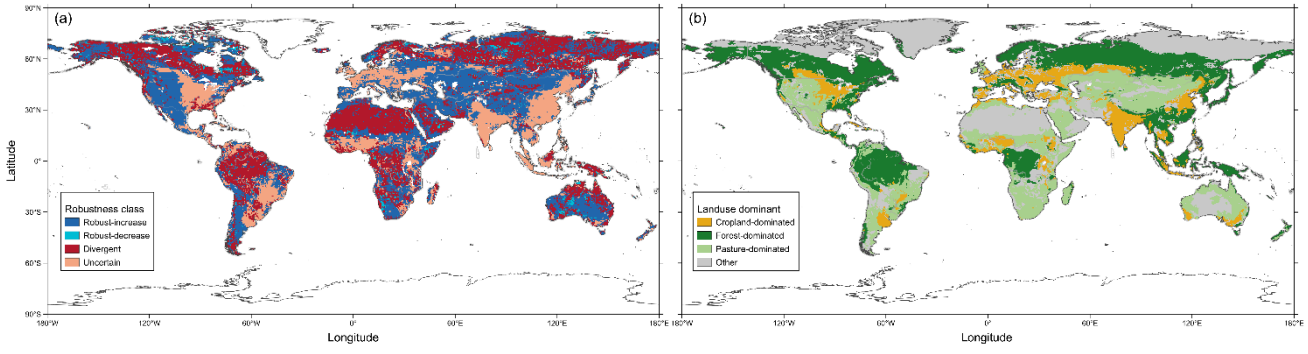

Figure 2. Global maps of (a) robustness classification and (b) dominant land-use type at 0.5° resolution. In panel (a), each grid cell is classified into one of four robustness categories based on the diagnostic framework (Fig. 1): Robust-increase (dark blue), Robust-decrease (cyan), Divergent (red) and Uncertain (orange). In panel (b), dominant land-use types are defined from LUH2 land-use fractions (Hurtt et al., 2020) for the year 2015: Cropland-dominated (cropland fraction ≥ 0.3, orange), Forest-dominated (forest fraction ≥ 0.5, dark green), Pasture-dominated (pasture fraction ≥ 0.3, light green) and Other (grey).

The ensemble-mean emissions for the 2010s were highest in intensively managed agricultural regions such as eastern China, the Indo-Gangetic Plain and western Europe (Fig. S1), and in most land areas emissions had increased since the 1850s, with the largest increases concentrated in the same regions (Fig. S2). The four classes exhibited distinct spatial patterns. Robust-increase grid cells were concentrated in mid-latitude pasture and rangeland regions, including central Eurasia, sub-Saharan Africa and southern South America (Fig. 2a). Robust-decrease cells were sparse and scattered, with small clusters in southeastern Australia. Divergent cells were most prevalent in forest-dominated regions (44.2% of all Divergent grid cells), particularly in the Amazon basin, the Congo basin and Southeast Asia, where the agreement κ_A declined to 0.500−0.625 (Fig. S3). A substantial proportion (41.8%) also occurred in regions classified as Other (neither cropland-, forest-, nor pasture-dominated), including arid and high-latitude zones, where both ensemble-mean 2010s emissions (median 0.009 vs. 0.039 g N m^{-2} yr^{-1} for forests) and inter-model spread (SD(Δ) median 0.003 vs. 0.016) were small in absolute terms, and directional disagreement occurred among individually small fluxes. Uncertain cells were concentrated in major agricultural regions, including eastern China, the Indo-Gangetic Plain, western Europe and the US Corn Belt.

Comparison of the robustness map (Fig. 2a) with the dominant land-use map (Fig. 2b) revealed a striking visual correspondence: Divergent cells (red) largely overlapped with forest-dominated regions (dark green), while Uncertain cells

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

(orange) coincided with cropland-dominated regions (orange). This spatial correspondence generated the land-use stratification presented in Section 3.2. These land-use dependent patterns were also evident at the regional scale (Fig. S5). Regions with intensive agriculture, such as South Asia (64.7% Uncertain) and China (45.2% Uncertain), were dominated by quantitative uncertainty, while tropical forest regions, such as Equatorial Africa and the Amazon Basin, had the highest Divergent proportions. Central Asia, dominated by pasture and rangeland, had the highest proportion of Robust-increase (85.6%).

3.2 Land-use dependence of model uncertainty

3.2.1 Agreement and inter-model spread

Cropland-dominated grid cells (n = 8,203) exhibited the highest agreement, with κ_A concentrated at 0.875−1.000 (Fig. 3a). The median κ_A for croplands reached 1.000, indicating that virtually all models agreed on the direction of N_2O change. However, croplands also exhibited the largest inter-model spread: the median SD(Δ) was 0.072 g N m^{-2} yr^{-1}, far exceeding that of forest (0.019) and pasture (0.017) (Fig. 3b).

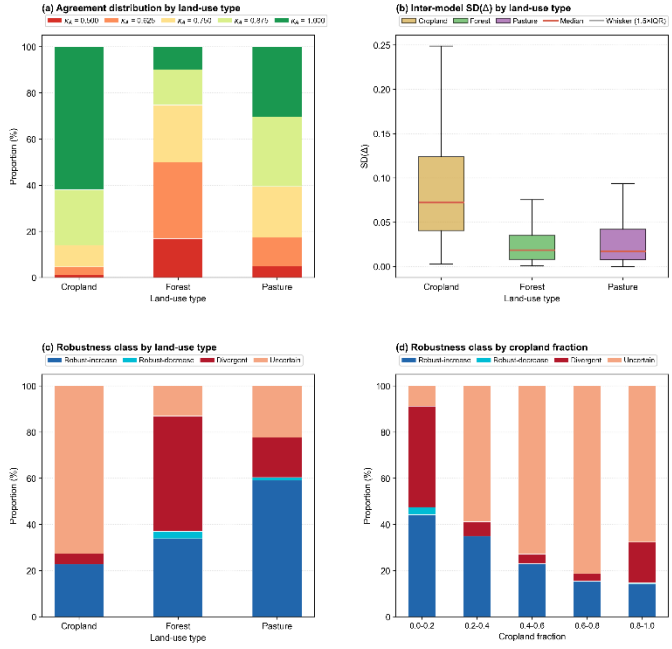

Figure 3. Comparison of robustness diagnostics across land-use types. (a) Proportional distribution of model agreement (κ_A) for cropland-, forest-, and pasture-dominated grid cells, shown as the proportion of cells at each agreement level (κ_A = 0.500 to 1.000). (b) Box plots of the inter-model standard deviation of change (SD(Δ)) by land-use type, with boxes indicating the interquartile range and horizontal lines indicating the median. (c) The proportion of robustness classes (Robust-increase, Robust-decrease, Divergent and Uncertain) by land-use type. (d) The proportion of robustness classes as a fraction of cropland fraction, binned at 0.2 intervals, showing the progressive

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

shift from Divergent-dominated to Uncertain-dominated uncertainty as the cropland fraction increased. Land-use classifications follow the thresholds defined in Section 2.4.

Forest-dominated grid cells (n = 18,175) exhibited the opposite pattern: agreement was low, with κ_A distributed across 0.500−0.750, while SD(Δ) was comparatively small.

Pasture-dominated grid cells (n = 16,003) combined a relatively high agreement with low inter-model spread, yielding the most robust results among the three land-use types.

3.2.2 Robustness class composition

These contrasting agreement-spread profiles translated into distinct robustness class compositions (Fig. 3c, Table 1):

- Cropland: Dominated by Uncertain classification; Divergent and Robust-decrease classes were virtually absent.
- Forest: Dominated by Divergent classification; Uncertain and Robust-increase classes were secondary.
- Pasture: Dominated by Robust-increase classification; Uncertain and Divergent classes occurred at moderate levels.

These results demonstrated that the nature of model uncertainty differed fundamentally across land-use types: quantitative in croplands, directional in forests, and largely resolved in pastures.

Table 1. Robustness class composition (%) by land-use type.

Class	Cropland n = 8,203	Forest n = 18,175	Pasture n = 16,003
Robust-increase	22.7	33.8	59.1
Robust-decrease	0.0	3.1	1.2
Divergent	4.6	50.0	17.4
Uncertain	72.6	13.1	22.3

3.3 Relationship between nitrogen input and inter-model uncertainty

The relationship between total anthropogenic nitrogen input (HaNi) and inter-model uncertainty reinforces the land-use findings (Fig. 4). When all valid grid cells were stratified into quintiles by nitrogen input intensity, both the magnitude and the nature of uncertainty shifted systematically.

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

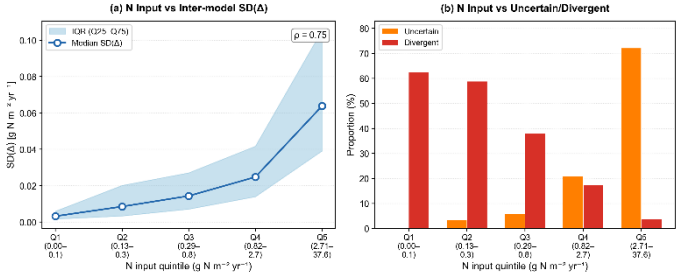

Figure 4. Relationship between anthropogenic nitrogen input and inter-model spread across nitrogen input quintiles. (a) Median inter-model standard deviation of change (SD(Δ); open circles) and interquartile range (IQR; shaded area) for each quintile. The Spearman rank correlation coefficient (ρ = 0.75) indicates a strong positive association between nitrogen input intensity and inter-model disagreement in the magnitude of change. (b) The proportion of grid cells classified as Uncertain (orange) and Divergent (red) within each quintile. Nitrogen input quintile boundaries (g N m^{-2} yr^{-1}) are shown in parentheses along the x-axis. Quintile ranges were derived from the HaNi dataset (Tian et al., 2022).

The median SD(Δ) increases approximately 20-fold from Q1 (0.0032 g N m^{-2} yr^{-1}) to Q5 (0.0638 g N m^{-2} yr^{-1}) (Fig. 4a). The Spearman rank correlation between nitrogen input and SD(Δ) was ρ = 0.75, confirming a strong monotonic association. The spatial pattern of SD(Δ) closely mirrored the distribution of high anthropogenic nitrogen inputs, with the largest inter-model spread concentrated in eastern China, South Asia and the US Corn Belt (Fig. S4).

Concurrently, the class composition shifted from Divergent-dominated to Uncertain-dominated (Fig. 4b): Divergent decreased from 62.5% (Q1) to 3.7% (Q5), while Uncertain increased from 0.1% (Q1) to 72.2% (Q5).

This pattern demonstrated that nitrogen input intensity governed not only the magnitude of inter-model spread but also a qualitative transition in the nature of uncertainty from directional disagreement (Divergent) under low nitrogen inputs to quantitative disagreement (Uncertain) under high nitrogen inputs. This transition was also visible along the cropland fraction gradient (Fig. 3d), where Divergent declined and Uncertain increased monotonically from low to high cropland fractions, which was consistent with the strong association between cropland coverage and nitrogen input intensity.

4 Discussion

4.1 Novelty of the framework

The RDF proposed here extends conventional approaches to multi-model ensemble assessment comprehensively and clearly. In previous NMIP studies, inter-model uncertainty has been simply characterized using the ensemble mean ± standard deviation (Tian et al., 2020, 2024). While informative, this approach treats uncertainty as a single dimension and cannot distinguish between regions where models disagree on the direction of change and regions where models agree on the direction but differ on the magnitude.

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

By integrating agreement (κ_A) alongside inter-model spread (SD(Δ)), our RDF separated these two qualitatively different types of uncertainty and systematically provided more in-depth insights. This distinction proved critical in the present analysis. Cropland-dominated regions and forest-dominated regions exhibited comparable overall uncertainty levels when measured by SD(Δ) alone, but the underlying structure of that uncertainty was fundamentally different: quantitative in croplands (Uncertain, 72.6%) versus directional in forests (Divergent, 50.0%).

The RDF builds on concepts developed for climate model ensembles. Tebaldi et al. (2011) proposed quantifying grid-cell-level model agreement for CMIP (Coupled Model Intercomparison Project) projections, and Knutti and Sedláček (2013) introduced the stippling/hatching convention combining signal-to-noise ratio and model agreement adopted in IPCC AR5 (Collins et al., 2013). Our approach adapted these ideas to the biogeochemical modelling context, focusing on historical changes and differences among land-use types. The resulting four-class system offered a more actionable diagnosis than continuous uncertainty metrics, with each class implying a distinct model improvement strategy (Section 4.5).

4.2 Drivers of quantitative uncertainty in cropland

The dominance of the Uncertain classification in cropland (72.6%) indicated that all eight models agreed on the direction of the N_2O increase since the 1850s. This was consistent with the well-established role of synthetic nitrogen fertilizer as the primary driver of agricultural N_2O emissions (Tian et al., 2020, 2024), but the magnitude of the increase differed substantially. Several factors may have contributed to this quantitative spread. First, the nonlinear response of N_2O emissions to nitrogen input is well documented from field observations (Shcherbak et al., 2014; Wang et al., 2020), and the degree to which each model captured this nonlinearity differed. The strong positive correlation between nitrogen input and SD(Δ) (Spearman ρ = 0.75) supported the interpretation that high-nitrogen conditions amplify inter-model differences in emission sensitivity. When the inter-model spread was normalized by the ensemble-mean change magnitude (CV(Δ)/$|\overline{\Delta}|$), the pattern was reversed. The median CV(Δ) decreased monotonically from Q1 (2.53) to Q5 (0.86), with a Spearman correlation of ρ = -0.60 ($p < 0.001$). Therefore, in high-nitrogen regions, the spread was large in absolute terms but small relative to the signal, with models agreeing not only on the direction but also on the relative magnitude of the N_2O increase. The absolute spread was therefore reflective of the large signal rather than the proportionally greater model divergence, supporting the interpretation that quantitative uncertainty in croplands is driven by genuine differences in emission sensitivity rather than by random noise.

Second, the eight NMIP2 models differed in their parameterization of nitrification and denitrification processes (Tian et al., 2024, Table S1). While all models represented these processes as functions of soil temperature and moisture, four models (ELM, O-CN, ORCHIDEE, and VISIT) additionally included soil pH dependence, which can substantially affect emission rates in the acidic soils typical of intensively managed croplands. Furthermore, only five models accounted for manure nitrogen inputs, meaning that the effective nitrogen loading differed across models for the same grid cell.

Third, the large quantitative spread was consistent with the known variability of EFs themselves. The IPCC Tier 1 default EF for direct N_2O emissions is 1 %, with an uncertainty range of 0.3−3% (IPCC, 2019), implying a factor of 10 range in expected

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

emissions per unit nitrogen input. Cui et al. (2021) further demonstrated that EFs vary by two orders of magnitude across global croplands, driven primarily by climatic and edaphic factors rather than management practices alone. These observation-based maps showed that EFs were particularly high in humid subtropical regions with gleysols and acrisols, coinciding with the regions where our analysis identified high SD(Δ) and set an Uncertain classification. This qualitative spatial correspondence suggests that the environmental conditions that produce high and variable EFs in reality also produce large inter-model spread, because models parameterize the sensitivity of nitrification and denitrification to these conditions differently.

The quantitative uncertainty identified here thus reflected a genuine scientific challenge, i.e. the inherent difficulty of representing the complex, nonlinear, and spatially heterogeneous relationship between nitrogen inputs and N_2O emissions, rather than a deficiency specific to any individual model.

4.3 Drivers of directional divergence in forests

The prevalence of the Divergent classification in forests (50.0%) indicated that models disagree not merely on the magnitude but on the very direction of N_2O emission changes since the 1850s. This directional split likely reflected the competing influence of multiple environmental drivers on forest N_2O dynamics.

In forest ecosystems, where direct anthropogenic nitrogen inputs are minimal, N_2O emission changes are driven primarily by indirect factors. Increasing atmospheric CO_2 concentrations can stimulate plant nitrogen uptake, reducing soil mineral nitrogen availability and, thereby, suppressing nitrification and denitrification, a pathway that tends to decrease N_2O emissions (Phillips et al., 2001; Zaehle et al., 2011). Conversely, climate warming enhances microbial activity, accelerating nitrogen mineralization and N_2O production, a pathway that tends to increase emissions (Smith, 1997; Butterbach-Bahl et al., 2013). The net effect was dependent on the relative strength of these opposing pathways, which was model-dependent and could lead to divergent outcomes.

Tian et al. (2024) reported that CO_2 fertilization effects on soil N_2O differ in sign across NMIP2 models, with ELM and ISAM producing positive effects and other models producing negative effects. This was consistent with our finding that forest grid cells frequently exhibited $\kappa_A = 0.500-0.625$, corresponding to a near even split among models. The lower SD(Δ) in forests compared to croplands (median 0.019 vs. 0.072 g N m^{-2} yr^{-1}) further supported this interpretation. The disagreement is primarily about direction rather than magnitude, because the absolute changes in forest N_2O were small relative to croplands. Analogous directional disagreement among terrestrial biosphere models has been reported for soil organic carbon dynamics, where models diverge on the sign of century-long SOC (Soil Organic Carbon) changes due to differing representations of input, decomposition and their sensitivities to environmental drivers (Tian et al., 2015).

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

4.4 Robustness of pasture results

Pasture-dominated regions exhibited the highest proportion of Robust-increase (59.1%), combining relatively high agreement with low inter-model spread. Although pastures receive some anthropogenic nitrogen inputs through manure application and atmospheric deposition (Tian et al., 2022), these inputs are generally lower than in croplands, and their robustness classification contrasts sharply with that of forests.

This difference may be attributed to the relatively small aboveground biomass and simpler carbon-nitrogen cycling in grasslands compared to forests. In forests, the large standing biomass amplifies the CO_2 fertilization effect on plant nitrogen uptake, and the deep root systems and large soil organic matter pools amplify the sensitivity to warming-driven mineralization. In pastures, the competing CO_2 fertilization and warming effects that drive directional divergence in forests exert a smaller influence on N_2O dynamics, resulting in greater inter-model consistency.

This interpretation was exemplified by the SH1 experiment, which included all driving factors simultaneously. To confirm this hypothesis, factorial experiments separating CO_2 and climate effects are needed in future studies.

4.5 Implications for model improvement

The contrasting nature of uncertainty across different land-use types implied that they required different model improvement strategies. For croplands, the observational benchmarking of EFs is of high priority, particularly in high-nitrogen regions where the quantitative spread is largest. For forests, an improved mechanistic understanding of the competition between CO_2 fertilization and warming effects on soil nitrogen cycling is needed to resolve the directional disagreement. Pasture representations already achieve reasonable inter-model consistency in both direction and magnitude. The regional-scale results (Fig. S5) reinforce these model improvement priorities: the high prevalence of the Uncertain classification in South Asia and China, the world's two largest agricultural N_2O source regions, highlights the policy relevance of reducing quantitative model spread in these areas.

For croplands, priority regions included eastern China, the Indo-Gangetic Plain and the US Corn Belt, where nitrogen inputs were highest and the Uncertain classification was most prevalent. Expanding field measurement networks to monitor N_2O fluxes in these regions would provide the observational constraints needed to reduce quantitative spread across models.

For forests, the Amazon basin, the Congo basin and Southeast Asian tropical forests were prioritized regions where directional divergence occurred intensively. Long-term manipulation experiments and eddy covariance measurements in these ecosystems would help resolve the CO_2 fertilization-versus-warming balance that underlies the current directional disagreement.

4.6 Limitations and future perspectives

This study had several limitations that should be acknowledged. First, for clarity and simplicity, our analysis focused on inter-model consistency and did not include comparisons with observational data. Benchmarking which models are closer to reality would require spatially explicit N_2O flux observations, which remain sparse at the global scale.

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

Second, the robustness classification was dependent on the choice of thresholds for κ_A and SD, because we adapted data-driven thresholds (κ_A = 0.75 and SD_{Q75}). To confirm the robustness of the main findings, i.e. cropland Uncertain dominance and forest Divergent dominance, we confirmed that similar results were obtained using alternative threshold values (Table S2). Third, this study used cropland fractions derived from the LUH2 dataset for simplicity and did not account for crop types, management practices, or fertilizer application timing, which are known to influence N_2O emissions (Cui et al., 2021; Li et al., 2024).

Fourth, the present analysis is based on the SH1 experiment, which includes the combined effects of all driving factors. If individual fractional experiments (e.g. CO_2-only, climate-only, N-input-only) were available for spatial analysis, it would be possible to attribute the inter-model differences to specific driving factors and provide more mechanistic insights. Such process-level decomposition represents a natural and valuable extension of the present framework.

Finally, the RDF itself was general and not specific to N_2O or NMIP. It could be applied to any multi-model ensemble where grid-cell-level change estimates are available, including TRENDY (carbon cycle), ISIMIP (impacts), and other MIP initiatives. The specific thresholds for agreement and inter-model spread adapted here were tailored to the NMIP2 ensemble and should be re-evaluated through a sensitivity analysis for each new application. Users should also note that the resolution of the agreement metric κ_A was dependent on the number of participating models. With eight models, κ_A took five discrete values (0.500 to 1.000), which was sufficient for the four-class classification, but ensembles with fewer than five models would yield very coarse agreement estimates and may require alternative approaches.

Looking ahead, the application of this RDF to future NMIP phases with enhanced agricultural metadata and factorial experiment outputs would substantially improve diagnostics. In particular, crop-type and management-specific analyses would allow the quantitative uncertainty identified in croplands to be decomposed into its component sources, while fractional decomposition would clarify the relative roles of CO_2, climate, and nitrogen inputs in driving the directional divergence observed in forests.

5 Conclusions

We proposed a robustness diagnostic framework that integrates change magnitude, model agreement, and inter-model spread into a four-class classification for multi-model ensembles. The RDF's key innovation was the explicit separation of directional disagreement (Divergent) from quantitative disagreement (Uncertain), a distinction that is not captured by the conventional ensemble mean ± standard deviation approach.

The application of the RDF to NMIP2 soil N_2O emissions (1850s vs. 2010s) from eight terrestrial biosphere models yielded the following findings:

1. The nature of model uncertainty differed fundamentally across land-use types. Cropland-dominated regions were dominated by Uncertain (72.6%). The models agreed that N_2O emissions have increased, but the magnitude of the increase varied substantially across models. Forest-dominated regions were dominated by Divergent (50.0%). The

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

models disagreed on whether N_2O emissions have increased or decreased. Pasture-dominated regions were the most robust, with 59.1% classified as Robust-increase.

2. Inter-model spread increased strongly with nitrogen input intensity (Spearman $\rho = 0.75$), and a systematic transition from Divergent-dominated to Uncertain-dominated uncertainty occurred along the cropland fraction gradient. This confirmed that process representations under high-nitrogen conditions were a key source of quantitative uncertainty.
3. These contrasting patterns implied differentiated model improvement strategies: observational benchmarking of EFs for croplands, particularly in high-nitrogen regions such as eastern China, the Indo-Gangetic Plain, and the US Corn Belt; and improved mechanistic understanding of the competition between CO_2 fertilization and warming effects for forests, particularly in tropical regions.

The proposed RDF was applicable generally and was not specific to N_2O or NMIP. It could be applied to any multi-model ensemble where grid-cell-level change estimates were available, although the specific thresholds for agreement and inter-model spread should be re-evaluated through a sensitivity analysis for each application. Application of this RDF to future NMIP phases with enhanced agricultural metadata and factorial experiment outputs would enable a substantially more detailed diagnosis, moving from pattern identification to the mechanistic attribution of model disagreement.

Code and data availability

The NMIPs soil N_2O emission data are available from the NMIP2 data archive (Tian et al., 2024; https://doi.org/10.18160/RQ8P-2Z4R). The LUH2 land-use harmonization dataset is available from https://luh.umd.edu (Hurtt et al., 2020). The 0.5°-regridded HaNi nitrogen input data used in this study were provided by A. Ito (The University of Tokyo, Japan) for use in NMIP2 simulations. The original 5 arcmin HaNi dataset is publicly available from PANGAEA (Tian et al., 2022; https://doi.org/10.1594/PANGAEA.942069). Analysis code is available from Zenodo (Inatomi, 2026; https://doi.org/10.5281/zenodo.19083012).

Author contributions

Motoko Inatomi designed the study, performed the analysis and wrote the manuscript.

Competing interests

The author has declared that there are no competing interests.

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

Acknowledgements

The author thanks Akihiko Ito (The University of Tokyo, Japan) for providing the 0.5°-regridded HaNi nitrogen input data prepared for NMIP2 simulations.

An AI language model (Claude, Anthropic) was used to organize ideas and refine the manuscript structure iterative discussion. The authors reviewed, verified and take full responsibility for all content.

Financial support

This research was supported by institutional funding from the National Agriculture and Food Research Organization (NARO), Japan.

References

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?, Phil. Trans. R. Soc. B, 368, 20130122, 2013.

Collins, M., Knutti, R., Arblaster, J., et al.: Long-term Climate Change: Projections, Commitments and Irreversibility, in: Climate Change 2013: The Physical Science Basis, Cambridge University Press, 1029–1136, 2013.

Cui, X., Zhou, F., Ciais, P., et al.: Global mapping of crop-specific emission factors highlights hotspots of nitrous oxide mitigation, Nature Food, 2, 886–893, https://doi.org/10.1038/s43016-021-00384-9, 2021.

Davidson, E. A.: The contribution of manure and fertilizer nitrogen to atmospheric nitrous oxide since 1860, Nature Geosci., 2, 659–662, https://doi.org/10.1038/ngeo608, 2009.

Forster, P., Storelvmo, T., Armour, K., et al.: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity, in: Climate Change 2021: The Physical Science Basis, Cambridge University Press, 923–1054, https://doi.org/10.1017/9781009157896.009, 2021.

Hurtt, G. C., Chini, L., Frolking, S., et al.: Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6, Geosci. Model Dev., 13, 5425–5464, https://doi.org/10.5194/gmd-13-5425-2020, 2020.

Inatomi, M.: Analysis code for "A robustness diagnostic framework for NMIP ensembles: Application to NMIP2 soil N_2O emission estimates" (v1.0.0), Zenodo [code], https://doi.org/10.5281/zenodo.19083012, 2026.

IPCC: 2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories, Vol. 4, Ch. 11, IPCC, 2019.

Knutti, R. and Sedláček, J.: Robustness and uncertainties in the new CMIP5 climate model projections, Nature Clim. Change, 3, 369–373, https://doi.org/10.1038/nclimate1716, 2013.

Li, W., Tian, H., Pan, N., et al.: Machine learning emulation of NMIP2 process-based models to project global N_2O emission factors under climate change, Global Change Biol., 30, e17472, https://doi.org/10.1111/gcb.17472, 2024.

https://doi.org/10.5194/egusphere-2026-1817
Preprint. Discussion started: 11 May 2026

Phillips, R. L., Whalen, S. C., and Schlesinger, W. H.: Response of soil methanotrophic activity to carbon dioxide enrichment in a North Carolina coniferous forest, Soil Biol. Biochem., 33, 793–800, https://doi.org/10.1016/S0038-0717(00)00227-3, 2001.

Ravishankara, A. R., Daniel, J. S., and Portmann, R. W.: Nitrous oxide (N_2O): The dominant ozone-depleting substance emitted in the 21st century, Science, 326, 123–125, https://doi.org/10.1126/science.1176985, 2009.

Shcherbak, I., Millar, N., and Robertson, G. P.: Global meta analysis of the nonlinear response of soil nitrous oxide (N_2O) emissions to fertilizer nitrogen, P. Natl. Acad. Sci. USA, 111, 9199–9204, https://doi.org/10.1073/pnas.1322434111, 2014.

Smith, K. A.: The potential for feedback effects induced by global warming on emissions of nitrous oxide by soils, Global Change Biol., 3, 327–338, https://doi.org/10.1046/j.1365-2486.1997.00100.x, 1997.

Tebaldi, C., Arblaster, J. M., and Knutti, R.: Mapping model agreement on future climate projections, Geophys. Res. Lett., 38, L23701, https://doi.org/10.1029/2011GL049863, 2011.

Tian, H., Lu, C., Yang, J., Banger, K., Huntzinger, D. N., Schwalm, C. R., Michalak, A. M., Cook, R., Ciais, P., Hayes, D., Huang, M., Ito, A., Jain, A. K., Lei, H., Mao, J., Pan, S., Post, W. M., Peng, S., Poulter, B., Ren, W., Ricciuto, D., Schaefer, K., Shi, X., Tao, B., Wang, W., Wei, Y., Yang, Q., Zhang, B., and Zeng, N.: Global patterns and controls of soil organic carbon dynamics as simulated by multiple terrestrial biosphere models: Current status and future directions, Global Biogeochem. Cycles, 29, 775–792, https://doi.org/10.1002/2014GB005021, 2015.

Tian, H., Yang, J., Lu, C., et al.: The Global N_2O Model Intercomparison Project, B. Am. Meteorol. Soc., 99, 1231–1251, https://doi.org/10.1175/BAMS-D-17-0212.1, 2018.

Tian, H., Xu, R., Canadell, J. G., et al.: A comprehensive quantification of global nitrous oxide sources and sinks, Nature, 586, 248–256, https://doi.org/10.1038/s41586-020-2780-0, 2020.

Tian, H., Xu, R., and Canadell, J. G., et al.: HaNi: a historical dataset of anthropogenic nitrogen inputs to the terrestrial biosphere 1860–2019, Earth Syst. Sci. Data, 14, 4551–4568, https://doi.org/10.5194/essd-14-4551-2022, 2022.

Tian, H., Pan, N., Thompson, R. L., et al.: Global nitrous oxide budget (1980–2020), Earth Syst. Sci. Data, 16, 2543–2604, https://doi.org/10.5194/essd-16-2543-2024, 2024.

Wang, Q., Zhou, F., Shang, Z., et al.: Data-driven estimates of global nitrous oxide emissions from croplands, Natl. Sci. Rev., 7, 441–452, 2020.

World Meteorological Organization (WMO): WMO Greenhouse Gas Bulletin No. 19: The State of Greenhouse Gases in the Atmosphere Based on Global Observations through 2022, WMO, Geneva, 15 Nov 2023. (Available at: https://wmo.int/publication-series/wmo-greenhouse-gas-bulletin-no-19)

Zaehle, S., Ciais, P., Friend, A. D., and Prieur, V.: Carbon benefits of anthropogenic reactive nitrogen offset by nitrous oxide emissions, Nature Geosci., 4, 601–605, https://doi.org/10.1038/ngeo1207, 2011.