Reply to RC2: Anonymous Referee #2

Manuscript: egusphere-2026-1817

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

Author: Motoko Inatomi

Date: 29 July 2026

We sincerely thank Referee #2 for taking the time to read our manuscript carefully and for providing thoughtful and constructive comments. The Referee's insights have helped us to improve the clarity and scientific rigour of the manuscript.

To facilitate a constructive exchange, we would like to briefly summarize the conceptual basis of our study. The RDF proposed in this study is designed to diagnose the nature of inter-model uncertainty — specifically, whether models disagree in the direction (Divergent) or magnitude (Uncertain) of N_2O change — at the grid-cell level, using publicly available ensemble outputs. This diagnostic approach is complementary to, and distinct from, mechanistic attribution studies that require factorial experiment outputs or process-level decomposition. We have carefully considered each comment in this context and provide our detailed responses below.

Comment 1:

Title: This framework is general and the NMIP2 N_2O emission is only an application case. Therefore, I would suggest to remove 'for NMIP ensembles' or replace it as 'for multiple-model ensembles'.

Response:

We thank the Referee for this suggestion. We agree that the RDF is general and not specific to NMIP ensembles. We will revise the title to "A robustness diagnostic framework for multiple-model ensembles: Application to NMIP2 soil N_2O emission estimates."

Comment 2:

As all of the analysis is based on the N_2O differences between 1850s and 2010s, for consistency, I would suggest to also use the N inputs differences rather than the absolute N inputs. The N fertilizer is zero in 1850s, but not for N manure and N deposition.

Response:

We thank the Referee for this comment. We would like to clarify the rationale for using absolute N input values. The central question in Section 3.3 is whether regions with currently high nitrogen loading show greater inter-model uncertainty in N_2O change. To address this question, absolute N input values for 2015 are the appropriate metric, as they represent present-day agricultural management intensity. Using ΔN inputs would instead address a fundamentally different question — whether regions where nitrogen application has increased since the 1850s show greater inter-model uncertainty — which is not the objective of our analysis. The choice of metric was therefore guided by

the scientific question being addressed rather than by the form of the response variable. We therefore believe that the use of absolute N input values is the most appropriate choice for addressing the scientific question posed in Section 3.3.

Comment 3:

Sect. 3.3: I don't think this session is informative. Because the high N inputs usually happen in cropland while the low N inputs happen in forest, this analysis basically repeat the results that N_2O in cropland consistently increase but varied in magnitude, while the forest N_2O emissions divergent in trend directions. Why not separate the fertilizer, manure and N deposition and get the similar figure individually?

Response:

We thank the Referee for this comment. We would like to clarify the distinction between the objective of Section 3.3 and the mechanistic attribution approach suggested by the Referee.

	Mechanistic attribution	**Robustness diagnostic (this study)**
Question	Why do models disagree?	How do models disagree?
Objective	Identify drivers of inter-model differences	Diagnose the nature of inter-model uncertainty
Required data	Factorial experiment outputs	Ensemble outputs only
Output	Attribution of uncertainty to specific drivers	Classification of uncertainty as Divergent or Uncertain

Section 3.3 examines whether the total nitrogen loading is associated with the nature of inter-model uncertainty (Divergent vs Uncertain), not why that uncertainty arises. Separating N inputs into fertilizer, manure, and deposition would address a mechanistic attribution question, which is beyond the scope of the present study. We hope this clarification helps to convey the intended purpose of Section 3.3, and we believe that the current analysis is well suited to the diagnostic objectives of this study. We acknowledge that mechanistic attribution — addressing *why* models disagree — is an equally important dimension, and we encourage future studies to pursue such analyses.

Comment 4:

Continued on Fig.4, instead of plotting the absolute SD variations with N inputs, why not using emission factor? As the author has already discussed, the absolute SD is expected to increase with higher N inputs, but whether the nonlinear responses of EF are well captured by models remain unknown. It is good to know model divergences in not only N_2O emissions, but also the EF.

Response:

We thank the Referee for this interesting suggestion. We would like to clarify that the objective of Figure 4 is to examine whether the nature and magnitude of inter-model uncertainty in N_2O change varies with nitrogen loading, consistent with the diagnostic objectives of this study. The analysis of emission factor (EF) variability across models would address a different question — namely, whether models capture the nonlinear response of EF to nitrogen inputs — which falls within the mechanistic attribution dimension (Why) outlined in our conceptual framework. Furthermore, EF calculation requires both SH1 and SH3 experiment outputs (i.e., simulations with and without N fertilizer input; Tian et al., 2024), and the SH3 spatial outputs are not included in the NMIP2 public data archive. EF

analysis is therefore beyond the scope of the present study. We note that attribution of N_2O changes to specific driving factors has been conducted at global budget level by Tian et al. (2024) using NMIP2 factorial experiment outputs. However, spatially explicit mechanistic attribution of inter-model differences at the grid-cell level has not yet been conducted. We acknowledge that such grid-cell-level mechanistic attribution represents an important and necessary complement to the present diagnostic framework, and we encourage future studies to pursue this direction when factorial experiment outputs become available at the spatial level.

Comment 5:

Line 270: please use lowercase for 'Uncertainty'. Also Line 341

Response:

We thank the Referee for this comment. We note that "Uncertain" in Lines 270 and 341 refers to a specific robustness class defined within our diagnostic framework, and is capitalized throughout the manuscript as a proper class name, consistent with the treatment of the other class names (Robust-increase, Robust-decrease, and Divergent). We therefore respectfully maintain the capitalization.

Comment 6:

Lines 277: Confusing sentence. What pattern is reserved?

Response:

We thank the Referee for this comment. "The pattern was reversed" in Line 277 of the preprint referred to the reversal of the relationship between inter-model spread and nitrogen input intensity: whereas SD(Δ) increased monotonically with nitrogen input (Spearman $\rho = 0.75$), CV(Δ) decreased monotonically ($\rho = -0.60$). This sentence has been restructured in the revised manuscript to make this distinction explicit.