Reply to RC1: Anonymous Referee #1

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 thank the referee for their careful and constructive review of our manuscript. The comments have helped us to clarify the novelty and scope of the proposed framework. We respond to each comment below.

Major Comment 1:

The central methodological contribution of the manuscript is the RDF classification scheme, which combines three metrics: ensemble-mean change, model agreement on the sign of change, and inter-model standard deviation. These are all standard diagnostics in multi-model ensemble analysis. The main novelty appears to be the grouping of these diagnostics into four qualitative classes. Although this is useful for interpretation, I am not convinced that it constitutes a sufficiently substantial methodological advance for GMD in its present form. The manuscript should more clearly demonstrate what is new relative to existing robustness and model-agreement approaches, especially those already used in climate-model ensemble studies. A more convincing contribution would require a systematic comparison between the proposed RDF and existing diagnostics, such as sign agreement maps or robustness maps. The current manuscript reads more as an application of existing ensemble diagnostics to NMIP2 outputs than as a new model assessment method.

Response:

We appreciate this substantive concern and agree that the individual metrics used in the RDF — ensemble-mean change, sign agreement, and inter-model standard deviation — are established diagnostics in climate model ensemble analysis. We do not claim otherwise. However, we respectfully argue that the novelty of this study lies in three distinct contributions that go beyond the grouping of existing metrics.

First, the explicit separation of directional disagreement (Divergent) from quantitative disagreement (Uncertain) is absent in previous NMIP studies, where inter-model uncertainty has been characterized solely as ensemble mean ± standard deviation (Tian et al., 2020, 2024). This one-dimensional treatment of uncertainty cannot distinguish regions where models disagree on the direction of change from regions where models agree on direction but disagree on magnitude. As demonstrated in our analysis, this distinction is not merely taxonomic: cropland grid cells and forest grid cells exhibit comparable SD(Δ) values, yet their uncertainty structures are fundamentally different — quantitative in croplands (Uncertain, 78.2%) versus directional in forests (Divergent, 50.0%). This difference has direct implications for model improvement strategies, which cannot be identified by SD alone.

Second, this is the first application of a robustness diagnostic framework to a biogeochemical model intercomparison project focused on N_2O. While robustness maps and sign agreement maps have been applied to climate variables in CMIP (e.g., Tebaldi et al., 2011; Knutti and Sedláček, 2013), no equivalent framework has been applied to NMIP or to terrestrial N_2O flux ensembles. The transfer of this diagnostic culture to biogeochemical ensembles, where the nature of uncertainty differs fundamentally from temperature or precipitation projections due to the dominant role of nitrogen input intensity and land-use type, represents a meaningful contribution to the field.

Third, the RDF is designed to operate on ensemble outputs without requiring factorial experiment data. This makes it immediately applicable to existing and future NMIP phases, as well as to other biogeochemical model intercomparison projects.

We acknowledge the referee's suggestion to provide a systematic comparison between the RDF and existing diagnostics such as sign agreement maps. We note that the spatial distribution of model agreement (κ_A) is already presented in Fig. S3 of the Supplement. In the manuscript, we will explicitly direct readers to Fig. S3 in the main text and strengthen the discussion in Section 4.1 to clearly illustrate where and why the RDF classification provides additional insights beyond a conventional sign agreement map. We will also make the three points of novelty outlined above more explicit in Section 4.1.

Major Comment 2:

The definition of "Uncertain" mechanically classifies high-N-input regions as uncertain. The definition of the Uncertain class is based on sign agreement combined with an absolute inter-model standard deviation threshold. This design makes grid cells with high N input rates much more likely to be classified as Uncertain, because high N inputs naturally lead to larger absolute N_2O emissions and therefore larger absolute inter-model spread. As a result, almost all major agricultural regions, including eastern China, India, western Europe, the US Corn Belt, and southeastern Brazil, were classified as Uncertain (Figure 2a).

The manuscript's Figure 3 shows that model agreement is very high for cropland-dominated grid cells, and the median κ_A for croplands reached 1.000. The classification as Uncertain therefore does not primarily reflect disagreement in the direction of N_2O change, but rather the use of an absolute SD threshold in regions where the value itself is large. This raises a fundamental question: does the Uncertain category identify genuinely low-confidence regions, or does it primarily identify high-emission/high-N-input regions with large absolute fluxes? The framework would be more robust if it considered relative uncertainty, signal-to-noise ratio, normalized SD, or uncertainty relative to the magnitude of the ensemble-mean change.

Response:

We thank the referee for this important observation. We agree that the use of an absolute SD threshold warrants careful interpretation, and we appreciate the opportunity to clarify this point.

The concern that the Uncertain classification mechanically identifies high-N-input regions is directly addressed in Section 4.2 of the manuscript, where we present an analysis of relative uncertainty using the coefficient of variation $CV(\Delta) = SD(\Delta) / |\bar{\Delta}|$. When inter-model spread is normalized by the ensemble-mean change magnitude, the pattern reverses: the median $CV(\Delta)$ decreases monotonically from Q1 (2.53) to Q5 (0.86) (Table 2), with a Spearman correlation of $\rho = -0.60$ ($p < 0.001$). This demonstrates that, in high-N-input regions, the spread is large in absolute terms but small relative to the signal. Models agree not only on the direction but also on the relative magnitude of N_2O increase. The Uncertain classification in croplands therefore reflects actual differences in emission sensitivity among models, rather than a methodological artifact of using an absolute threshold.

We acknowledge, however, that this analysis is currently embedded within the broader discussion of Section 4.2 and may not have been sufficiently prominent. In the revised manuscript, we will restructure Section 4.2 to present the CV(Δ) analysis at the outset, as a direct and explicit response to the question of whether the Uncertain category identifies genuinely uncertain regions. We will also add a reference to Table 2 at the point where the Uncertain classification is first introduced in Section 3, to ensure readers encounter this evidence at the appropriate point in the manuscript.

Major Comment 3:

The analysis of cropland-, forest-, and pasture-dominated regions is one of the central components of the manuscript. However, the current classification is based on grid-cell-level land-use fractions, with thresholds of cropland ≥ 0.3, forest ≥ 0.5, and pasture ≥ 0.3. However, most 0.5° grid cells are mixtures of different land use types. Under the current definition, a grid cell with cropland fraction slightly above 0.3 can be classified as cropland/pasture dominated even if another biome or land-use type occupies a larger fraction of the same grid cell. The terms "cropland-dominated" and "pasture-dominated" may be misleading. A more appropriate analysis would be to conduct the diagnostic at the biome or PFT level, if such outputs are available from NMIP2 models, or at least to use stricter land-use classification criteria.

Response:

We thank the referee for this careful observation. We agree that the mixed land-use nature of 0.5° grid cells is a genuine limitation, and we appreciate the opportunity to clarify our classification approach.

Following the referee's suggestion to use stricter land-use classification criteria, we have revised the main analysis to use a cropland fraction threshold of ≥ 0.5, consistent with the forest threshold (≥ 0.5). The pasture threshold has similarly been revised to ≥ 0.5 for consistency. This change reduces the cropland grid cells from n = 8,203 (≥ 0.3) to n = 4,099 (≥ 0.5), ensuring that classified grid cells are more clearly dominated by the respective land use. The main findings — cropland Uncertain dominance (78.2%) and forest Divergent dominance (50.0%) — remain robust under this stricter criterion, confirming that the qualitative contrast between cropland and forest uncertainty structures is not an artifact of the threshold choice.

To further demonstrate robustness, we have expanded the sensitivity analysis in Table S2 to include cropland thresholds of ≥ 0.3, ≥ 0.5 (main analysis), and ≥ 0.7, as well as pasture thresholds of ≥ 0.3 and ≥ 0.5 (main analysis). The results confirm that the main findings are robust across all threshold combinations.

Regarding the suggestion to conduct the analysis at the PFT level, we agree that this would provide more precise attribution of uncertainty to specific vegetation types. However, the NMIP2 public data archive provides grid-cell total N_2O flux outputs only; PFT-level outputs are not available in the public release. This is a genuine limitation of the present study, and we have stated it explicitly in the Limitations section of the revised manuscript.

We have also revised the terminology throughout the manuscript, replacing "cropland-dominated", "forest-dominated", and "pasture-dominated" with "cropland grid cells", "forest grid cells", and "pasture grid cells", respectively. In addition, Section 2.4 now explicitly defines these terms and clarifies that they refer to grid cells where the respective land-use fraction exceeds the defined threshold, not necessarily grid cells where that land use is the single largest fraction.

We also acknowledge that the terms "cropland-dominated" and "pasture-dominated" may be misleading, as the referee points out. In the revised manuscript, we will add a sentence in Section 2.4 clarifying that these terms refer to grid cells where the respective land-use fraction exceeds the defined threshold, not necessarily to grid cells where that land use is the single largest fraction.

Major Comment 4:

The discussion attributes cropland uncertainty mainly to differences in N-input sensitivity and forest divergence to the competing effects of CO_2 fertilization and warming. These interpretations are plausible, but they remain largely speculative in the current manuscript. NMIP2 includes a series of factorial experiments designed to isolate the effects of different drivers. These experiments should be used to quantify the contributions of different drivers to inter-model disagreement. This would substantially strengthen the manuscript and move it from a descriptive classification exercise toward a mechanistic model-diagnostic study. Without using these factorial simulations, the manuscript cannot robustly identify the drivers of uncertainty. The current SH1-only analysis can describe where models disagree, but it cannot adequately explain why they disagree.

Response:

We thank the referee for this suggestion. We agree that factorial experiments would provide valuable mechanistic insights into the uncertainty patterns identified by the RDF. However, the spatial outputs of the factorial experiments (e.g. CO_2-only, climate-only, and N-input-only simulations) are not publicly available in the NMIP2 data archive at the grid-cell level. The present study was therefore designed to provide a spatially explicit characterization of publicly available SH1 outputs — an objective that is independently meaningful and does not require factorial experiment data to be achieved. Identifying where models agree and disagree — and what nature of uncertainty dominates in different regions — is itself a scientifically meaningful objective, independent of mechanistic attribution. This diagnostic step is a necessary precursor to mechanistic decomposition: knowing where uncertainty is directional versus quantitative helps to prioritize which regions and processes warrant factorial analysis in future NMIP phases.

We acknowledge that the current SH1-only analysis can identify where models disagree but cannot robustly attribute why they disagree. This is a limitation of the present study, and we will state it more explicitly in the revised manuscript. Specifically, we will revise the Limitations section to clarify that the factorial experiment outputs are not publicly available at the grid-cell level, and that mechanistic attribution of inter-model differences to specific drivers therefore remains beyond the scope of the present analysis.

We also note that the interpretations offered in Sections 4.2 and 4.3 — attributing cropland uncertainty to differences in N-input sensitivity and forest divergence to competing CO_2 and warming effects — are explicitly framed as plausible hypotheses supported by published literature (Tian et al., 2024; Butterbach-Bahl et al., 2013), rather than as definitive causal attributions. More fundamentally, we wish to clarify that the primary objective inter-model disagreement, but to characterize the nature of that disagreement — whether it is directional or quantitative, and where it is spatially concentrated. This diagnostic objective is distinct from, and complementary to, mechanistic attribution.

We argue that characterizing the nature and spatial distribution of uncertainty is itself a scientifically sufficient and meaningful contribution, independent of mechanistic attribution. The RDF establishes what needs to be explained and where: cropland regions require better constraints on emission sensitivity, while forest regions require better process understanding of competing CO_2 and warming effects. This spatially explicit prioritization of model improvement targets is achievable from public SH1 outputs alone and provides a concrete foundation for designing factorial analyses in future NMIP phases.

Minor Comments

Line 28: “Foster”→”Forster”

We thank the Referee for catching this typographical error. “Foster et al. (2021)” in Line 28 should read “Forster et al. (2021)”, consistent with the reference list. This will be corrected in the revised manuscript.

Line 74-75: grammatically incorrect sentence.

We thank the Referee for this comment. We agree that the sentence is grammatically awkward. The phrase “common driving data” is intended to indicate that all models shared the same nitrogen input datasets. We will revise the sentence to read: “… mineral nitrogen fertilization and manure nitrogen inputs as common driving data.” This will be clarified in the revised manuscript.

Lines 154–155: clarify how overlapping grid cells are handled

We thank the Referee for raising this point. The handling of overlapping grid cells is already described in Section 2.4 (Lines 153-154): “These categories are not mutually exclusive; overlapping grid cells are retained in both categories.” For Figure 2b, “dominant land-use type” refers to the category with the highest fractional coverage in that grid cell, as defined in the figure caption (Lines 173-175). To avoid confusion, we will add a cross-reference in the figure caption to Section 2.4, while the statistical stratification in Section 3.2 uses the fractional thresholds described therein.

Figure 3: panels are small and difficult to read

We acknowledge that the individual panels in Figure 3 are small. In the revised manuscript, we will increase the figure dimensions to span the full page width and enlarge the font sizes to improve legibility of labels, legends, and axis text.

Line 350, 362: "adapted" → "adopted"

We thank the Referee for this correction. “Adapted” should read “adopted” at both Line 350 and Line 362, as the thresholds were adopted from data-driven criteria without modification. Both will be corrected in the revised manuscript.

Line 397: "NMIPs" → "NMIP2"

We thank the Referee for this correction. “NMIPs” in Line 397 should read “NMIP2”, as the analysis is based specifically on the second phase of NMIP. This will be corrected in the revised manuscript.