This manuscript egusphere-2026-1094 presents a multi-model ensemble framework combining three process-based agroecosystem models (DAYCENT, DNDC, ECOSYS) to simulate corn and soybean yields as well as soil organic carbon (SOC) stock changes across the continental United States (CONUS) from 2014–2023. The study is spatially comprehensive (4 km² resolution) and leverages harmonized environmental, soil, and management datasets. The ensemble median generally outperforms individual models when compared against NASS yield data and RaCA SOC measurements, showing reduced RMSE and improved central tendency. The authors conclude that ensemble modeling reduces structural uncertainty and provides a more robust basis for carbon accounting and sustainable agricultural policy. While the topic is timely and relevant to SOIL, the manuscript contains several conceptual, methodological, and presentational issues that need substantial revision before publication. Please find my comments below.

Author's response: We would like to thank the reviewer for their time and thoughtful critique of this manuscript. We have addressed all the comments and believe that the article has been improved because of the valuable feedback. Please find our point-by-point responses below.

Major comments:

1. The manuscript oscillates between claiming the ensemble "represents" uncertainty (e.g., lines 38–39, Page 3) and "reduces" uncertainty (e.g., lines 17–18, Page 2). These are different scientific goals. Please clarify: is the ensemble intended to characterize the range of plausible outcomes from structural variability, or to improve predictive accuracy via error cancellation? The framing affects the interpretation of RMSE improvements and the value of including biased models like ECOSYS.

Author's response: We appreciate the reviewer's request for clarification. We have revised the manuscript (specifically in the abstract, introduction & conclusion section) to consistently frame that the ensemble framework can be used as a tool to characterize structural uncertainty to capture the range of plausible outcomes resulting from the different modeling assumptions inherent in DAYCENT, DNDC, and ECOSYS. While the ensemble average is used as an estimator of the mean condition (which empirically improves predictive accuracy when compared to NASS and RaCA data), we clarify that the primary value of the multi-model approach lies in its ability to quantify the spread of model responses across diverse CONUS environments. We acknowledge that including a model with different sensitivities (eg., ECOSYS) is not about correcting the mean, but about ensuring the structural variability of agroecosystem processes is fully represented in our SOC and yield prediction.

2. ECOSYS consistently underperforms for both corn and SOC. The manuscript acknowledges this but does not convincingly justify why ECOSYS should be retained in the ensemble beyond "complementary strengths." If ECOSYS is mechanistically more detailed but poorly calibrated for CONUS, its inclusion may degrade rather than

enhance ensemble performance. Please provide a clearer rationale, or consider a sensitivity analysis excluding ECOSYS.

Author's response: We agree with the reviewer that ECOSYS's performance metrics are lower than those of DAYCENT and DNDC in this specific application. However, we have chosen to retain ECOSYS to ensure that the ensemble framework captures a biologically plausible range of structural uncertainties. In response to the comment, we have revised the limitations section to provide a clearer rationale. We now explicitly discuss how ECOSYS's mechanistic complexity makes it more sensitive to the uniform phenotypic parameters (such as maturity groups) used for computational efficiency at the CONUS scale. We characterize this as a sensitivity penalty where a high-complexity model is constrained by simplified regional inputs rather than a structural failure. By utilizing the ensemble mean, we mitigate the impact of this underperformance on our central estimates while allowing the inter-model spread to provide a more conservative quantification of the uncertainty inherent in agroecosystem modeling. We believe that excluding ECOSYS would result in an artificially narrow uncertainty range that overlooks known mechanistic sensitivities.

3. Section 2.3 uses only RMSE for model evaluation and parameter fine-tuning. RMSE does not penalize model complexity, nor does it account for systematic bias or pattern similarity. Please add additional metrics (e.g., Nash-Sutcliffe efficiency, percent bias, or Akaike Information Criterion as you suggested) to better characterize model performance. Also clarify whether the same data were used for calibration and validation to avoid overfitting.

Author's response: We thank the reviewer for this important clarification and completely agree that the distinction between calibration and independent validation is critical for interpreting model performance.

To evaluate the potential for model overfitting and assess framework stability across both SOC and crop yields, we tested a stratified data-splitting protocol using the createDataPartition framework from the caret package. This algorithm protects against sampling bias in continuous datasets by first dividing the data (RaCA SOC or NASS yields) into equal-sized numeric bins based on quantiles, and then randomly sampling exactly 50% of the data points from each individual bin to form the calibration set, leaving the remaining 50% for independent validation. This stratification ensures that both subsets mirror the exact same underlying mean, variance, and outlier distribution. This approach resulted into nearly identical error distributions and performance metrics between the independent groups: an Ensemble RMSE of 4.7 kg/m2 (calibration) vs. 4.6 kg/m2 (validation) for soil organic carbon, 3.0 Mg/ha (calibration) vs. 3.0 Mg/ha (validation) for corn yield, and 1.0 Mg/ha (calibration) vs. 1.0 Mg/ha (validation) for soybean yield. Because of this identical performance explicitly shows that modeling framework is structurally stable and free from overfitting, we opted to pool the subsets and present the full dataset in the final manuscript figures to fully utilize the available datasets and maintain geographic continuity across the CONUS. We intentionally opted not to report R2 or Nash-Sutcliffe Efficiency (NSE), as these metrics can be misleadingly inflated in regional contexts where models are closely aligned with the regional observation mean; instead, we rely on RMSE

and normalized RMSE (%RMSE) to provide an interpretable measure of error magnitude across the CONUS. To further enhance spatial transparency, we have incorporated additional analyses in the supplementary material, including maps showing the spatial distribution of the RMSE of yield and SOC datasets (Figures S1 and S2) show structural uncertainty, providing a more rigorous assessment of the framework's current capabilities while acknowledging that independent validation on entirely external temporal datasets remains a priority for future research.

4. In the method section (Section 2.3), it appears that the same datasets (NASS crop yield data and RaCA SOC data) have been used for both model parameterization (i.e., fine-tuning to minimize RMSE) and subsequent model validation. This practice risks overfitting and can lead to overly optimistic performance metrics, including the reported RMSE improvements for the ensemble. Please clarify whether any form of data separation (e.g., temporal or spatial holdout, cross-validation) was applied. If not, the reported agreement between simulated and observed values may reflect calibration rather than true predictive skill, and the ensemble's apparent advantage could be overstated.

Author's Response: We thank the reviewer for this important clarification. We acknowledge the concern regarding the potential for overfitting when utilizing the same datasets for both model parameterization and performance evaluation.

To rigorously test for regional sampling bias and evaluate the potential for model overfitting across the CONUS, we executed a diagnostic stratified data-splitting protocol using the createDataPartition framework. This algorithm protects against sampling imbalances in continuous datasets by dividing the complete range of empirical observations (RaCA SOC or NASS yields) into equal-sized numeric bins based on quantiles. It then randomly samples exactly 50% of the data points from each individual bin to form a calibration set, reserving the remaining 50% as an independent validation holdout. This stratification ensures that both subsets share identical statistical properties (mean, variance, and outlier distributions).

As detailed in our response to Comment 3, data split yielded nearly identical error distributions and performance metrics between the independent calibration and validation groups. As RMSE values were mathematically indistinguishable, it explicitly demonstrated that our regional multi-model framework is structurally stable and entirely free from over-calibration. Consequently, to fully utilize all datasets, and maintain complete geographic continuity across the CONUS, we opted to pool the data for the final manuscript figures.

We have revised Section 2.3 to explicitly detail this stratified split diagnostic and clarify that the final metrics represent a globally optimized calibration across the full domain, ensuring each model architecture is evaluated at its best achievable regional fit. We contend that the ensemble's advantage remains highly significant: even across the pooled domain, the ensemble mean consistently smooths out individual model discrepancies. This structural diversity effectively mitigates the systematic biases of any single model architecture, providing a robust, transparent

baseline of current process-based modeling capabilities at this scale while independent validation on entirely external temporal datasets remains a priority for future research.

5. The manuscript states that "model parameterization was conducted to minimize RMSE between observed and predicted values" but does not describe the spatial/temporal splitting of NASS and RaCA data. If the same years and locations used for calibration are also used for evaluation, the reported RMSE values likely underestimate true prediction uncertainty. Please discuss how this might affect the apparent ensemble improvement.

Author's Response: We agree with the reviewer that because the models are evaluated across the same spatial and temporal domains used for calibration, the reported RMSE reflects the error bound of a globally optimized regional calibration rather than a fully independent prediction uncertainty.

As detailed in our responses to Comments 3 and 4, we explicitly evaluated this potential underestimation of uncertainty by executing a diagnostic stratified split using the createDataPartition framework. This technique uses 50% independent spatial holdout across identical quantile bins to mirror the full dataset's distribution. As shown by the metrics reported in our response to Comment 3, the ensemble error bounds were identical between the calibration and independent validation subsets across all soil carbon and crop yield. This mathematical convergence directly demonstrates that the apparent ensemble improvement is not due to localized overfitting or over-calibration. While we acknowledge that a fully independent temporal holdout would capture a broader range of true predictive uncertainty which remains a priority for future work as more data becomes available, the relative advantage of the ensemble mean over individual models remains a highly robust finding of this study. We have expanded our discussion in the revised manuscript to explicitly contextualize these error values as an optimized baseline, outlining limitation of our current uncertainty estimates and the clear mathematical advantage of the multi-model ensemble.

6. The finding that the Midwest and Southeast show SOC gains while the Great Plains and West show losses under corn-soybean rotation is important but under-discussed. Why would the same rotation cause SOC losses in drier/western regions? Is this due to lower baseline SOC, different decomposition rates, or management differences? Similarly, DNDC projected SOC losses exceeding 0.01 kg C m^{-2} yr^{-1} in high-yielding zones. Please expand the mechanistic interpretation.

Author's response: We thank the reviewer for this insightful comment. We agree that the divergent SOC trends between the humid regions (Midwest and Southeast) and the semi-arid regions (Great Plains and West) need more detailed mechanistic explanation in the manuscript.

The observed SOC losses in the Great Plains and West under corn-soybean rotations are primarily driven by a negative carbon mass balance. In these regions, water-limited Net Primary Productivity (NPP) results in lower biomass returns, while periodic moisture and high temperatures maintain high rates of heterotrophic respiration. Furthermore, the low-residue incorporation due to low yield in these drier environments often fails to meet the minimum carbon input required to maintain SOC equilibrium.

Regarding the high-yielding zones, DNDC projections suggest that the combination of nitrogen availability and optimal soil moisture and temperature profiles create a high carbon turnover. Here, accelerated microbial activity can mineralize existing soil organic matter faster than new residues are humified. We have expanded the discussion to incorporate these biogeochemical mechanisms. Note that DNDC use Double Monod kinetic equation where microbial growth is very sensitive to availability of C and N (Li et al., 1992), whereas model like DAYCENT use conservative first-order pool based approach where N primarily limits plant productivity rather than acting as a direct catalyst for the mineralization of stabilized SOM.

7. The caption and title of Figure 6 currently read "ECOSYS model projected SOC change…" but the figure shows (and text describes) multiple models and the ensemble. This must be corrected to "Agroecosystem model projected…" or similar.

Author's response: We thank the reviewer for identifying this oversight. We have corrected the title and caption of Figure 6 as suggested

8. You state that the ensemble "captures uncertainty", but with only three models, it is unlikely that their spread represents the full range of uncertainty across all existing agroecosystem models. Please add a discussion of how the three-model spread compares to published multi-model ensemble studies (e.g., AgMIP) and whether the results are robust to inclusion of other models (e.g., APSIM, EPIC).

Aurthor's response: We thank the reviewer for this constructive comment. We acknowledge that a three-model ensemble does not represent the range of structural uncertainty found in larger initiatives like AgMIP. The selection of Ecosys, Daycent, and DNDC for this study was driven by their widespread use in regional carbon assessments and their direct alignment with our project objectives. While these three models capture distinct biogeochemical processes ranging from pool-based to microbial-kinetic approaches we agree that the inclusion of models such as APSIM or EPIC would likely widen the ensemble spread. We have added a statement to the method and limitations section to further clarify the choice made in this study.

Minor comments:

1. Line 13–14 (Page 2): Add units to RMSE values
Aurthor's response: Thank you for the comment, we have added unit in updated manuscript.

2. Lines 17-18 (Page 4): The claim of "actionable information for policymakers" is

unsupported by the current results (baseline only, no management comparisons). Please tone down or reframe.

Aurthor's response: We acknowledge the reviewer's comment that without explicit management scenario comparisons, actionable solution for specific policy interventions is limited. We have updated this statement to emphasize that the results provide a spatial baseline map and identify regional risk zones for SOC decline.

3. Lines 24 (Page 4): Change "projections" to "estimates" or "simulations" since the study does not forecast future conditions.
Aurthor's response: Thank you for the comment, we have updated the text in revised manuscript.

4. Line 27 (Page 5): Typo: "where grown" → "were grown."
Aurthor's response: Thank you for finding the typo, we have updated the text in revised manuscript.

5. Section 2.2 (Page 5–6): The model descriptions are dense and lack comparative synthesis. Please add a summary table or paragraph comparing key differences: process formulation (mechanistic vs. semi-empirical), required inputs, treatment of belowground processes, and yield determination logic. Also explicitly state why these three models were selected over others.
Aurthor's response: We thank the reviewer for the suggestion to provide a more comparative synthesis. We have added a rationale for our model selection to the methods section and expanded the limitations section to address the scope of our ensemble. Additionally, we have included a summary table (Table 1) that synthesizes the key structural differences between the models, specifically focusing on process formulations, belowground dynamics, and yield determination logic. These changes clarified why these three specific models were chosen to represent a broad spectrum of biogeochemical modeling approaches.

Table 1 Summary of biogeochemical process representations and model structures for the three-model ensemble.

	DAYCENT	DNDC	ECOSYS
Conceptual Approach	Semi-empirical / Pool-based	Mechanistic / Microbial-kinetic	Mechanistic / Thermodynamic
Yield	Potential biomass limited by moisture and Nitrogen (N)	Photosynthesis-driven; N and water restricted	Full canopy-root coupling; hourly CO2 fixation
Belowground Process	Empirical root distribution by soil depth	Dynamic root growth based on C-allocation	Detailed rhizosphere; ion diffusion and mass flow

SOC Turnover	First-order pool based (Active, Slow, Passive)	Double Monod kinetics; microbial biomass-driven	Microbial functional groups; substrate-specific decay

Note: Daycent = Daily Century; DNDC = Denitrification-Decomposition; Ecosys = Ecosystem model.

6. Figure 2: Use a color-blind friendly palette. Also explain why the ensemble shows many outliers despite lower RMSE.
Aurthor's response: Thank you for suggestion, we have updated figure 2 and figure 3 to color blind palette. Regarding the outliers, the ensemble achieves a lower RMSE by averaging out systematic biases, which mathematically tightens the interquartile range. This narrower distribution causes localized extreme values (outliers) in ensemble plot.

7. Line 17 (Page 7): Delete the extra "a" before "an RMSE of 2.7."
Aurthor's response: Thank you for finding the typo, we have updated the text in revised manuscript.

8. Line 43–44 (Page 8): You suggest ECOSYS bias may come from coupled root-canopy processes. Could you quantify/support this (e.g., from sensitivity analyses) using your model output?
Aurthor's response: We thank the reviewer for the suggestion to explain the ECOSYS bias. Based on our model outputs and previous sensitivity analyses, we have clarified that the underprediction is due to model's high sensitivity to nitrogen (N) availability and uptake kinetics. Specifically, ECOSYS employs a detailed rhizosphere sub-model where N-uptake is governed by ion diffusion and mass flow; in large-scale simulations with standardized N-input data, this can lead to localized N-stress that is more extreme than observation. We have added a sentence citing the specific sensitivity of ECOSYS to nitrogen-water coupling to support this claim.

9. Line numbers are renewed in every page.
Aurthor's response: Thank you for the comment, we have now updated it to continuous line number.

Citation: https://doi.org/10.5194/egusphere-2026-1094-RC2

Reference

Li, C., Frolking, S., and Frolking, T. A. (1992). A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity. *Journal of Geophysical Research: Atmospheres* **97**, 9759–9776.